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

Mingze Yuan¹,², Yingda Xia¹,*, Hexin Dong¹,², Zifan Chen², Jiawen Yao¹, Mingyan Qiu¹, Ke Yan¹, Xiaoli Yin³, Yu Shi³, Xin Chen¹, Zaiyi Liu⁴, Bin Dong², Jingren Zhou¹, Le Lu¹, Ling Zhang¹, Li Zhang²
¹DAMO Academy, Alibaba Group ²Peking University ³Shengjing Hospital ⁴Guangdong 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 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-
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,$$
\mathbf{G} = \{\mathbf{G}_i\}_{i=1}^{K},
$$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 $\sum_{i=1}^{K} \mathbf{G}_i = 1^{H \times W \times D}$. 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 $\mathbf{P} \in \mathbb{R}^{HW \times D \times C}$ 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,

$$
\mathbf{C} \leftarrow \mathbf{C} + \underset{N}{\arg \max} (\mathbf{Q}^c \times (\mathbf{K}^p)^T) \times \mathbf{V}^p,
$$

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 $\mathbf{P}^T$ 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,

$$
\mathbf{M} = \underset{N}{\text{softmax}}(\mathbf{R}) = \underset{N}{\text{softmax}}(\mathbf{C} \times \mathbf{P}^T).
$$

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,

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

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 $A_i = 1$ and $A_i = 0$ represent that $i$-th pixel in $\mathbf{X}$ belongs
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 \( Z \) and ground truth \( G \) in Eq. (1) as the segmentation loss, \( \ell_{\text{seg}} = \ell_{\text{ce}} + \ell_{\text{dc}} \). 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 \( G \in \mathbb{R}^{K\times HWD} \) to supervise the cluster assignment probability maps \( M \). 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 \( M \) (representing the assignment probabilities of the first \( N_1 \) cluster centers) with the background class \( G_1 \), the next \( N_2 \) channels with the organ class \( G_2 \), and the last \( N_3 \) channels with the tumor classes \( \sum_{i=3}^{K} G_i \). We define the merged cluster assignments \( \tilde{M} \) and class labels \( \tilde{G} \) as the following,

\[
\tilde{M} = (\tilde{M}_1, \tilde{M}_2, \tilde{M}_3) \in \mathbb{R}^{3\times HWD}
\]

\[
= \left( \sum_{i=1}^{N_1} M_i, \sum_{j=1}^{N_2} M_{N_1+j}, \sum_{k=1}^{N_3} M_{N_1+N_2+k} \right),
\]

(5)

\[
\tilde{G} = (\tilde{G}_1, \tilde{G}_2, \tilde{G}_3) = (G_1, G_2, \sum_{i=3}^{K} G_i) \in \mathbb{R}^{3\times HWD},
\]

(6)

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

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

\[
\ell_{\text{qd}} = - \sum_{j=1}^{HWD} \sum_{i=1}^{3} \tilde{G}_{ij} \log \tilde{M}_{ij},
\]

(7)

which draws strict boundaries between different types of cluster assignments (\( M_1, M_2, \) and \( M_3 \)) based on the ground truth. The final loss function \( L \) is a combination of segmentation loss \( \ell_{\text{seg}} \) and QD loss \( \ell_{\text{qd}} \) with a balance weight \( \lambda \), formulated as,

\[
L = \ell_{\text{seg}} + \lambda \ell_{\text{qd}}.
\]

### 3.3. Localizing OOD Regions with MaxQuery

Given a test image \( X \in \mathbb{R}^{H\times W\times D} \), our mask transformer will yield the pixel-level query response \( R \in \mathbb{R}^{N\times H\times W\times D} \), 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)
as the pixel-wise anomaly score, called MaxQuery, \( i.e., \)
\[
A = -\max_N R,
\]
where \( R \in \mathbb{R}^{N \times H \times W \times D} \) represents the query response matrix and \( A \in \mathbb{R}^{H \times W \times D} \) 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 \( R \) (pre-softmax, \( A = -\max_N R \)) and cluster assignments \( M \) (post-softmax, \( A' = -\max_N M \)). \( A \) greatly outperforms \( A' \), since if an inlier pixel is evenly close to multiple cluster centers, the maximal score in \( M \) can be very low and easily be mis-classified as an outlier. But with maximum query response \( R \) (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 oral (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 an attention layer in the block has 8 heads. Since medical image segmentation is sensitive to local textures, we add an attention layer in the block has 8 heads. Since medical image segmentation is sensitive to local textures, we add an attention layer in the block has 8 heads. Since medical image segmentation is sensitive to local textures, we add an attention layer in the block has 8 heads. Since medical image segmentation is sensitive to local textures, we add an attention layer in the block has 8 heads. Since medical image segmentation is sensitive to local textures, we add an attention layer in the block has 8 heads.

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
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 KMaxDeepLab [56] to directly add deep supervision on the attention map of \((k\text{-means})\) cross attention to align it with the final segmentation after the segmentation output head. The loss weight \(\lambda\) 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 receiver-operating characteristic 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 (SwinUNETR [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.

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.

Table 2, our segmentation performance for inliers surpasses nnUNet by 6.30% in DSC. We find that Swin UNETR [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
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.

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 \((16, 4, 12)\).

<table>
<thead>
<tr>
<th>Query Dist. ((N_1, N_2, N_3))</th>
<th>AUROC↑</th>
<th>AUPR↑</th>
<th>FPR↓</th>
<th>DSC\text{inlier}↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>SML [26]</td>
<td>56.10</td>
<td>30.44</td>
<td>77.81</td>
<td>35.47</td>
</tr>
<tr>
<td>(8, 4, 20)</td>
<td>84.44</td>
<td>51.32</td>
<td>42.10</td>
<td>36.50</td>
</tr>
<tr>
<td>(8, 20, 4)</td>
<td>83.73</td>
<td>49.76</td>
<td>43.32</td>
<td>39.79</td>
</tr>
<tr>
<td>(16, 4, 12)</td>
<td>82.52</td>
<td>55.60</td>
<td>46.90</td>
<td>41.77</td>
</tr>
<tr>
<td>(20, 4, 8)</td>
<td>85.66</td>
<td>55.17</td>
<td>37.24</td>
<td>38.19</td>
</tr>
<tr>
<td>(24, 4, 4)</td>
<td>86.41</td>
<td>52.58</td>
<td>33.70</td>
<td>39.43</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Level</th>
<th>Softmax</th>
<th>AUROC↑</th>
<th>AUPR↑</th>
<th>FPR95↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>post</td>
<td>58.14</td>
<td>16.28</td>
<td>79.29</td>
<td></td>
</tr>
<tr>
<td>pre</td>
<td>52.70</td>
<td>24.59</td>
<td>88.40</td>
<td></td>
</tr>
<tr>
<td>Query</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>post</td>
<td>76.88</td>
<td>33.98</td>
<td>55.82</td>
<td></td>
</tr>
<tr>
<td>pre</td>
<td>82.52</td>
<td>55.60</td>
<td>46.19</td>
<td></td>
</tr>
</tbody>
</table>

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.

Pre-softmax versus Post-softmax for MaxQuery. As shown in Tab. 4, MaxQuery with pre-softmax score \(R\) exceeds the one with post-softmax \(M\) 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 MaxLogit (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. 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.

### Table 1. Dataset details of real-world Pancreatic Tumors.

<table>
<thead>
<tr>
<th>Abbr.</th>
<th>Full name</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDAC</td>
<td>Pancreatic ductal adenocarcinoma</td>
<td>366</td>
</tr>
<tr>
<td>IPMN</td>
<td>Intraductal papillary mucinous neoplasms</td>
<td>61</td>
</tr>
<tr>
<td>PNET</td>
<td>Pancreatic neuroendocrine tumor</td>
<td>35</td>
</tr>
<tr>
<td>SCN</td>
<td>Serous cystic neoplasms</td>
<td>46</td>
</tr>
<tr>
<td>CP</td>
<td>Chronic pancreatitis</td>
<td>43</td>
</tr>
<tr>
<td>SPT</td>
<td>Solid pseudopapillary tumor</td>
<td>32</td>
</tr>
<tr>
<td>MCN</td>
<td>Mucinous cystadenoma</td>
<td>7</td>
</tr>
<tr>
<td>AC</td>
<td>Ampullary cancer</td>
<td>46</td>
</tr>
<tr>
<td>DC</td>
<td>Bile duct cancer</td>
<td>12</td>
</tr>
<tr>
<td>“other”</td>
<td>Other rare neoplasms</td>
<td>13</td>
</tr>
</tbody>
</table>

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].

<table>
<thead>
<tr>
<th>Abbr.</th>
<th>Full name</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCC</td>
<td>Hepatocellular carcinoma</td>
<td>162</td>
</tr>
<tr>
<td>ICC</td>
<td>Intrahepatic cholangiocarcinoma</td>
<td>51</td>
</tr>
<tr>
<td>Meta.</td>
<td>Metastasis</td>
<td>97</td>
</tr>
<tr>
<td>Heman.</td>
<td>Hemangiomas</td>
<td>75</td>
</tr>
<tr>
<td>Cyst</td>
<td>Cyst</td>
<td>146</td>
</tr>
<tr>
<td>Hepato.</td>
<td>Hepatoblastoma</td>
<td>17</td>
</tr>
<tr>
<td>FNH</td>
<td>Focal nodular hyperplasia</td>
<td>27</td>
</tr>
<tr>
<td>“other”</td>
<td>Other rare tumors</td>
<td>60</td>
</tr>
</tbody>
</table>

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.

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

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

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>nnUNet [24]</td>
<td>57.22</td>
<td>28.16</td>
<td>52.81</td>
<td>77.55</td>
<td>46.49</td>
<td>52.45</td>
</tr>
<tr>
<td>(nn)TransUNet [7]</td>
<td>52.26</td>
<td>25.50</td>
<td>42.31</td>
<td>70.90</td>
<td>47.52</td>
<td>47.70</td>
</tr>
<tr>
<td>Ours</td>
<td>67.61</td>
<td>30.78</td>
<td>60.40</td>
<td>77.07</td>
<td>47.61</td>
<td>56.69</td>
</tr>
</tbody>
</table>
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


[14] Hexin Dong, Zifan Chen, Mingze Yuan, Yuting 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


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


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


[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


