A. Appendix

A.1. Dataset Details

We provide the abbreviation and full name for each disease from *Pancreatic Tumors* and *Liver Tumors* in Tabs. A1 and A2, 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 [2, 8]. 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 [2, 8] 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 [10] (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 A1. 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 [2, 8].

A.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. A1. 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.

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 A2. 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 [10].

A.3. Baselines for Inlier Segmentation

Comparison with Other Baselines. For a fair comparison with our method, we train UNet [7], UNet++ [12], TransUNet [1] based on the framework of nnUNet [5]. 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 Tab. A3, our method shows superiority on inlier segmentation compared with strong baselines, including nnUNet [5] and (nn)TransUNet [1]. This demonstrates that the distinctive architecture of our newly designed mask transformers leads to better performance on real-world medical image segmentation.

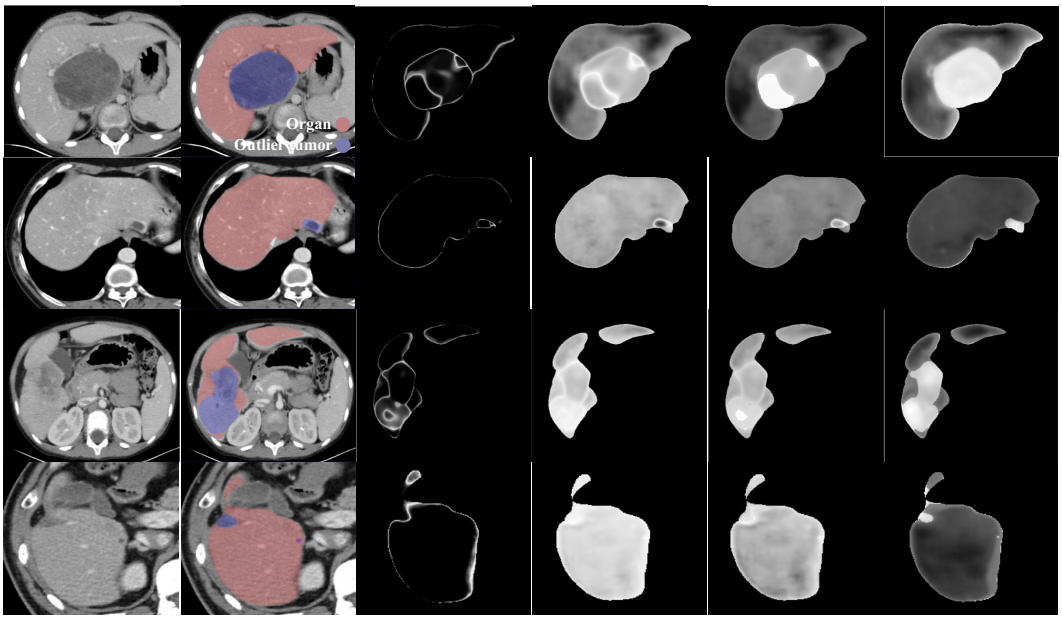
We also train Swin UNETR [9] using their officially released code and pre-trained model. We find that Swin UNETR [9] could not converge to reasonable tumor segmentations on *Pancreatic Tumors*, that might be due to its difficulty in identifying subtle tumor differences without sufficient data samples. Meanwhile, Swin UNETR [9] 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%.

A.4. Statistical Analysis

The Wilcoxon signed-rank test shows our method shows significant improvement to the second-best approaches on all metrics with $p < 0.01$, as presented in Tab. A4.

A.5. Hyper-parameter Selection.

We discuss in detail the key hyper-parameter of our method, i.e., (N_1, N_2, N_3), for controlling the query distribution, in Tab. 3 and Sec. 4.4. Our method shows robustness to different settings of query distribution. And another important hyper-parameter is the number of queries. It should be redundantly larger than the possible/useful classes in the data, which depends heavily on the data and the task. For other hyper-parameters on data augmenta-



(a) (b) (c) (d) (e) (f)

Figure A1. 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 [4], (d) MaxLogit [3], (e) SML [6] 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).

	Pancreatic %								Liver %					
Methods	PDAC	IPMN	PNET	SCN	CP	SPT	MCN	Avg.	HCC	ICC	Meta.	Heman.	Cyst	Avg.
UNet [7]	63.96	21.07	21.72	30.70	17.88	33.96	18.10	29.62	61.59	28.76	43.77	65.01	37.39	47.30
UNet++ [12]	63.43	22.85	14.52	25.09	15.02	21.36	10.07	24.62	56.51	29.13	36.88	56.74	46.60	45.17
TransUNet [1]	64.91	31.18	26.78	38.96	22.39	29.87	30.27	34.91	52.26	25.50	42.31	70.90	47.52	47.70
nnUNet [5]	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	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 A3. Inlier segmentation Dice scores (%) on *val* set of *Pancreatic Tumors* and *Liver Tumors* (all methods report results with final checkpoint). Our method notably outperforms all baselines for the task of inlier tumor segmentation.

p	AUROC	AUPR	FPR95	DSC
Pancreas	4.4×10^{-6}	2.0×10^{-6}	2.7×10^{-7}	2.0×10^{-6}
Liver	2.3×10^{-3}	7.0×10^{-3}	6.7×10^{-3}	2.8×10^{-3}

Table A4. Results of Wilcoxon signed-rank test versus the second-best approaches on all metrics.

tion, pre-processing, network architecture, and optimization, we follow the original settings in nnUNet [5] and KMax-Deeplab [11].

References

[1] 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. 1, 2

[2] 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. 1

[3] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *ICML*, 2022. 2

[4] 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. 2

[5] 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. 1, 2

[6] 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. 2

[7] 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. 1, 2

[8] 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. 1

[9] 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. 1

[10] 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. 1

[11] 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

[12] 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. 1, 2