上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

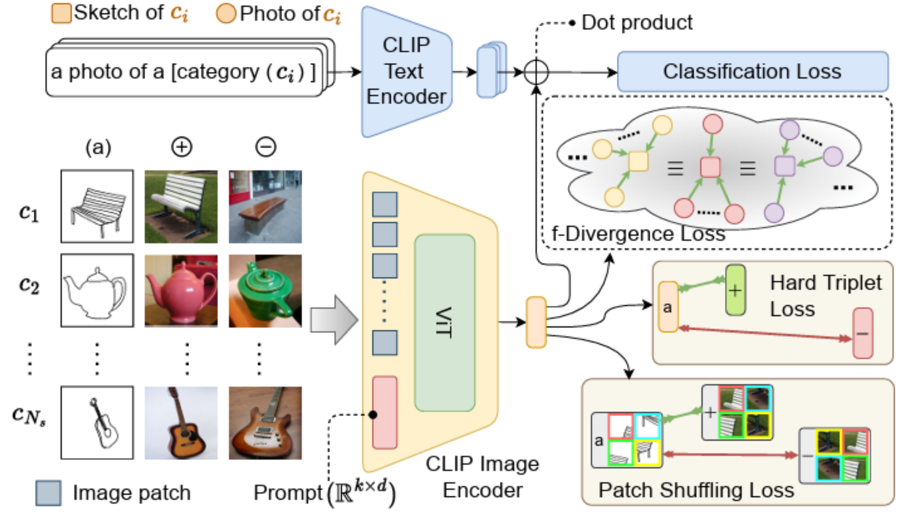
Unleashing the Potential of Vision-Language Pre-Training for 3D Zero-Shot Lesion Segmentation via Mask-Attribute Alignment

Yankai Jiang[1], Wenhui Lei[1,2], Xiaofan Zhang[1,2], Shaoting Zhang[1]

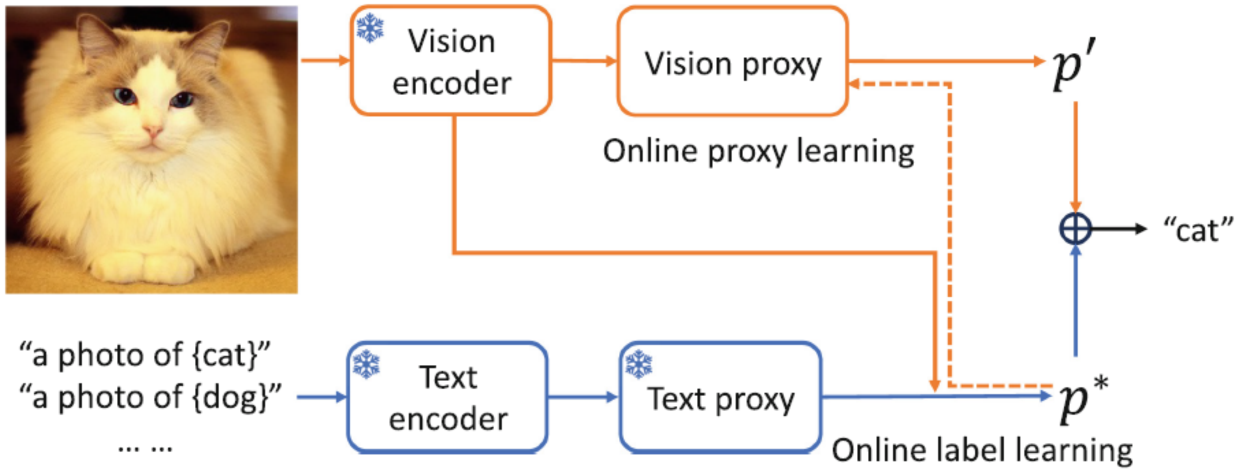
[1]Shanghai AI Laboratory [2]Shanghai Jiao Tong University

Background

Vision-language pre-training methods, e.g., CLIP, has illuminated a new paradigm for zero-shot object recognition.

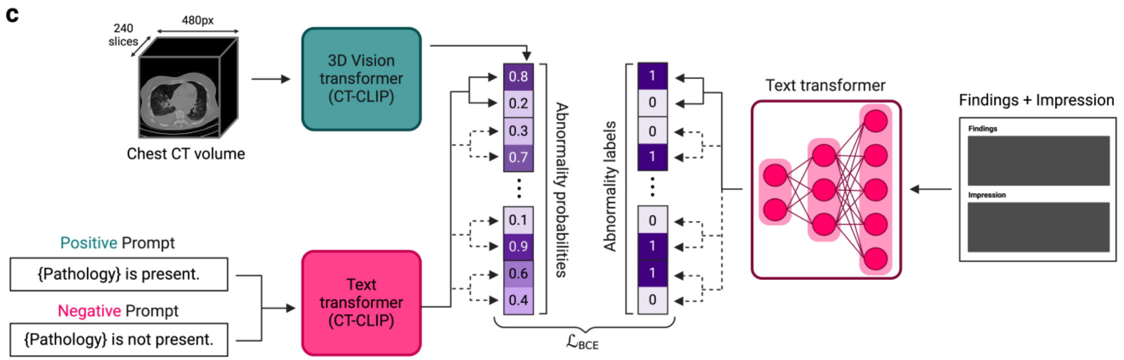


Sain, Aneeshan, et al. CVPR 2023

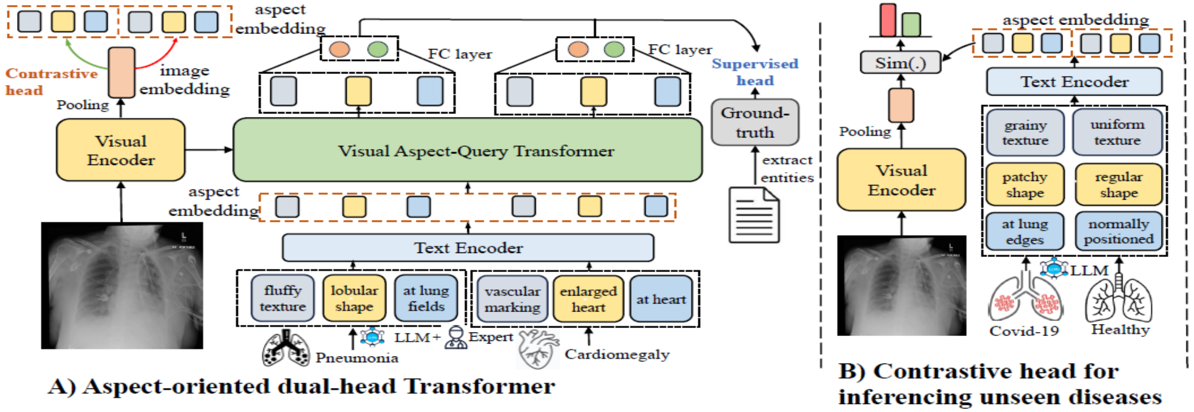


Qian, Qi, et al. ECCV 2024

This breakthrough also paves the way for significant advancements in zero-shot disease detection and diagnosis.



Hamamci, Ibrahim Ethem, et al. Arxiv 2024



Vu Minh Hieu Phan, et al. CVPR 2024

Motivation

Can we leverage the zero-shot capability of vision-language pre-training for 3D lesion segmentation?

Given the diversity and prevalence of new anomalies in clinical scenarios, along with the challenges of medical data collection, there is an increasing demand for zero-shot models capable of handling unseen diseases in an open-set setting.

Challenges:

- The substantial gap between the upstream contrastive pre-training task and the downstream per-pixel dense prediction task. The former focuses on aligning image-level global representations with text embeddings, while the latter requires fine-grained lesion-level visual understanding.
- Lesions can exhibit significant variations in shape and size, and present with blurred boundaries. Models struggle when encountering unseen lesion types due to their out-of-distribution visual characteristics. Simply using text inputs, such as raw reports, or common knowledge of disease definitions, is insufficient.

Key Ideas

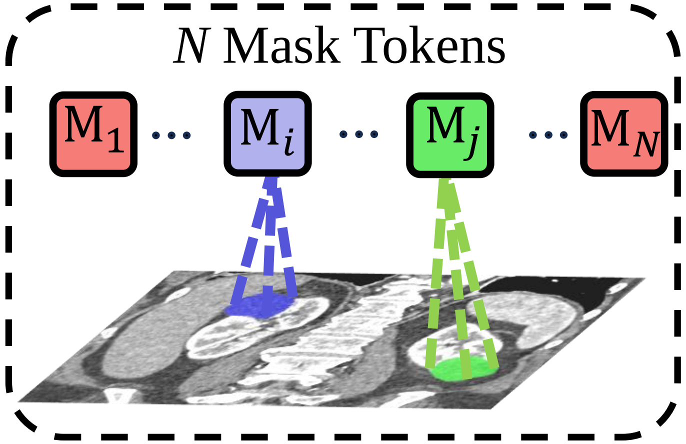
01 Leveraging multiscale mask representations with inherent boundary information to capture diverse lesion regions.

02 To learn extensible text representations that are robust to the out-of-distribution visual characteristics of unseen lesions, we incorporate domain knowledge from human experts to structure textual reports into descriptions of various elemental disease visual attributes (e.g., shape, intensity, location).

03 Multi-scale mask-attribute alignment aligns disease region features with different attributes, forming multiple positive pairs for each lesion mask to establish fine-grained relationships between visual features and various disease attributes.

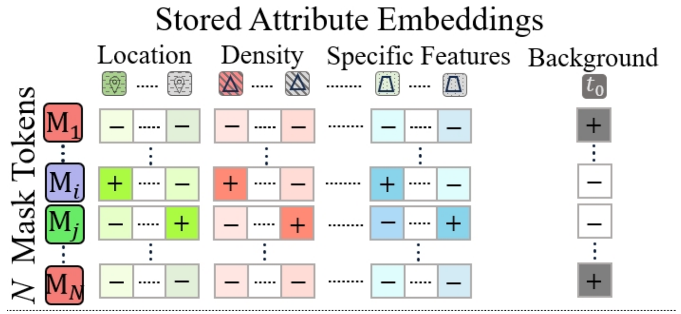
04 Cross-Modal Knowledge Injection (CMKI) module leverages both enhanced mask and attribute embeddings to generate predictions

01

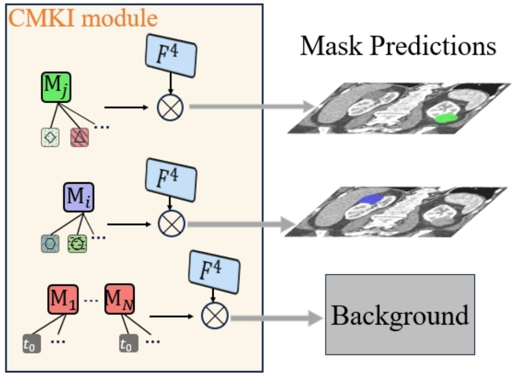


02

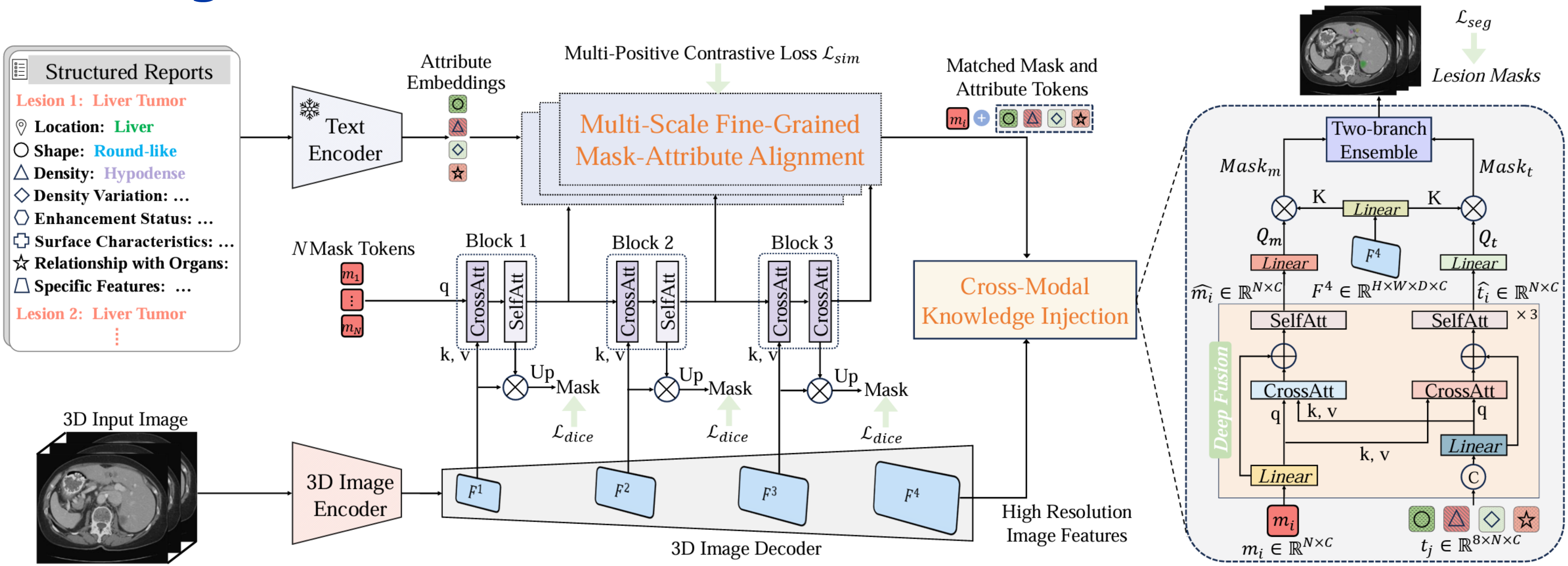
03



04

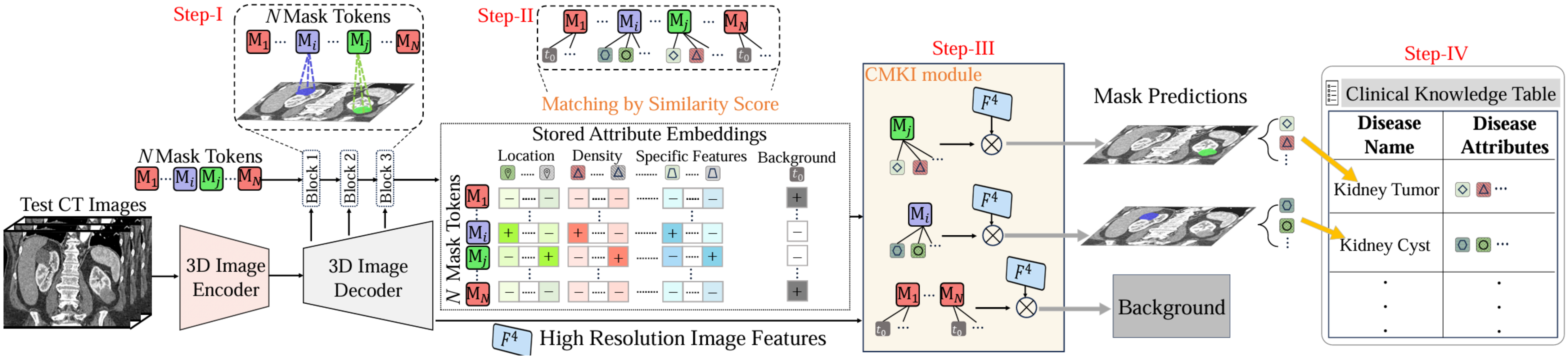


Training



- Utilization of Multi-Scale Features.
- Dissecting Reports into Descriptions of Fundamental Disease Attributes.
- Multi-Positive Contrastive Loss
- Cross-Modal Knowledge Injection

Testing



- Step-I: Image Partitioning via Mask Tokens. Test CT images are divided into regions, each represented by mask tokens.
- Step-II: Mask-attribute matching. Each mask token is associated with stored attribute embeddings.
- Step-III: Cross-modal fusion and mask prediction. Information from mask tokens and text embeddings is fused to generate segmentation masks.
- Step-IV: Disease identification via attribute-querying. The Clinical Knowledge Table links the predicted attributes to specific disease categories for precise diagnosis.

Results

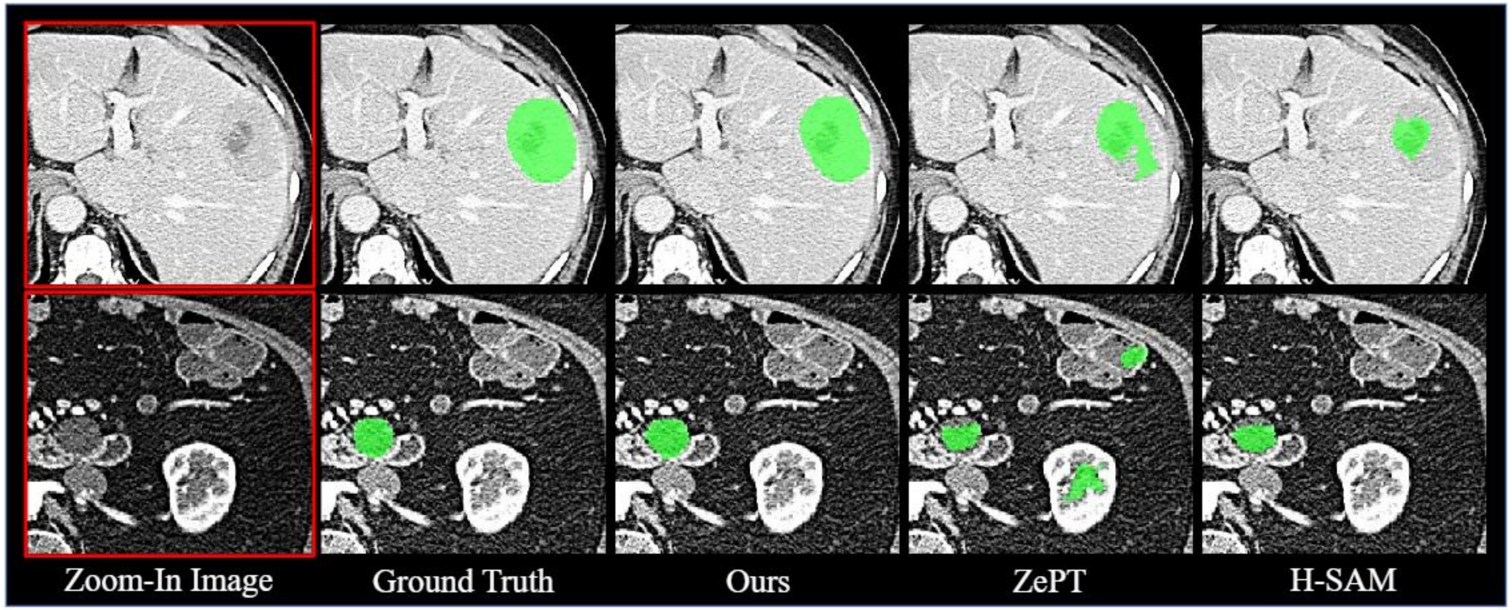
Segmentation Performance on Seen Classes

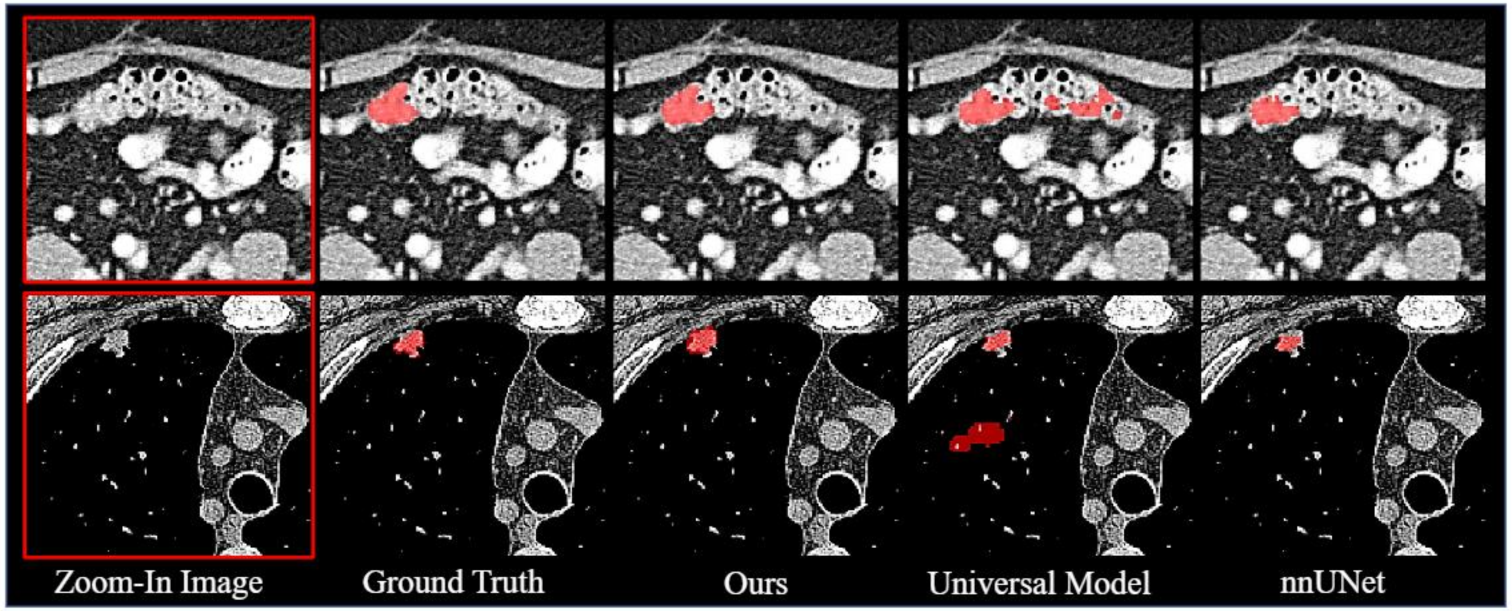
Method	MSD								KiTS23	
	Colon Tumor		Pancreas Tumor		Liver Tumor		Lung Tumor		Kidney Cyst	
	DSC↑	NSD↑	DSC↑	NSD↑	DSC↑	NSD↑	DSC↑	NSD↑	DSC↑	NSD↑
TransUNet∗	44.78±16.21	54.14±15.67	38.85±10.25	54.72±11.59	60.05±5.29	72.88±5.98	67.13±6.08	68.89±7.22	48.43±14.04	52.32±15.62
nnUNet∗	47.02±15.85	57.36±14.33	37.97±10.54	53.98±11.86	61.33±5.01	73.27±5.44	69.50±5.61	71.39±6.55	48.76±13.82	52.96±15.19
Swin UNETR∗	46.87±16.02	55.28±15.52	38.72±10.33	54.01±11.67	62.37±4.88	74.75±5.09	68.95±5.67	71.03±6.82	48.06±14.26	52.11±16.05
Universal Model∗	51.02±14.62	60.93±13.36	42.40±9.54	58.54±10.79	64.25±3.94	77.06±4.21	67.27±5.71	69.33±6.95	50.25±12.24	54.17±13.53
Malenia	**53.55±13.49**	**62.41±12.81**	**43.30±9.29**	**59.63±10.55**	**65.18±3.74**	**78.95±4.03**	**70.96±5.56**	**72.34±6.29**	**51.60±11.84**	**55.41±12.99**

Zero-shot Abilities.

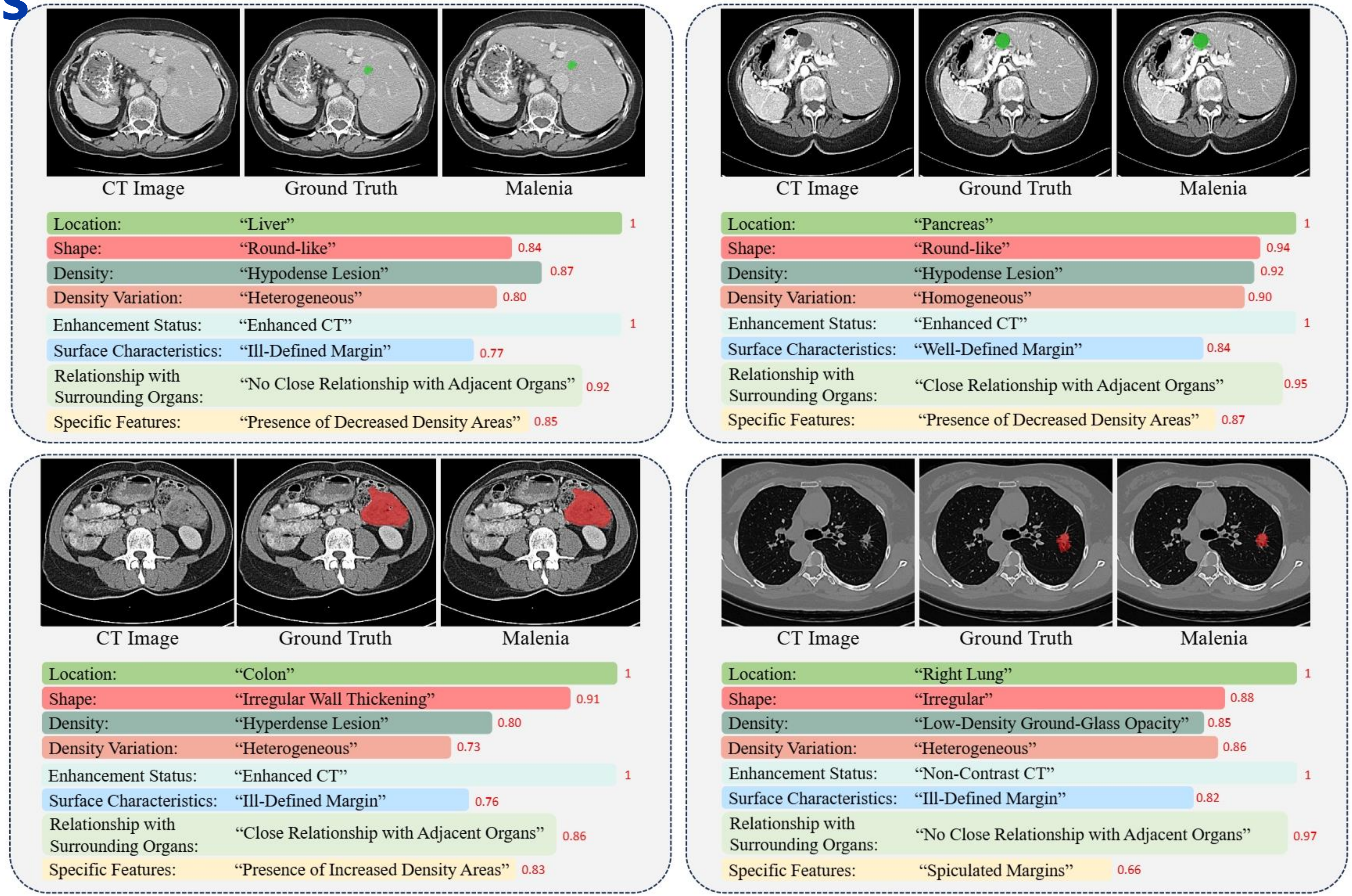
Method	MSD				KiTS23		In-house Dataset					
	Hepatic Vessel Tumor		Pancreas Cyst		Kidney Tumor		Liver Cyst		Kidney Stone		Gallbladder Tumor	
	DSC↑	NSD↑	DSC↑	NSD↑	DSC↑	NSD↑	DSC↑	NSD↑	DSC↑	NSD↑	DSC↑	NSD↑
SAM† (Shaharabany & Wolf, 2024)	35.76	45.83	37.17	49.26	35.45	41.33	34.99	40.88	24.14	31.92	28.08	36.38
SAM2† (Yamagishi et al., 2024)	35.93	45.88	38.42	50.85	35.67	41.88	35.29	41.25	25.50	33.74	28.57	36.62
SaLIP∗ (Aleem et al., 2024)	39.65	48.71	41.92	53.06	38.64	44.91	37.71	44.26	27.24	36.61	30.84	38.97
H-SAM* (Cheng et al., 2024)	45.58	54.24	46.87	57.91	44.21	50.39	43.75	50.20	29.23	38.11	32.17	40.05
ZePT∗ (Jiang et al., 2024)	53.12	63.25	53.35	63.50	46.82	52.44	51.64	57.36	33.97	42.42	35.48	43.23
Malenia	**59.52**	**69.60**	**60.91**	**70.28**	**54.96**	**60.60**	**61.85**	**70.93**	**43.05**	**52.95**	**47.35**	**55.79**

Qualitative visualizations

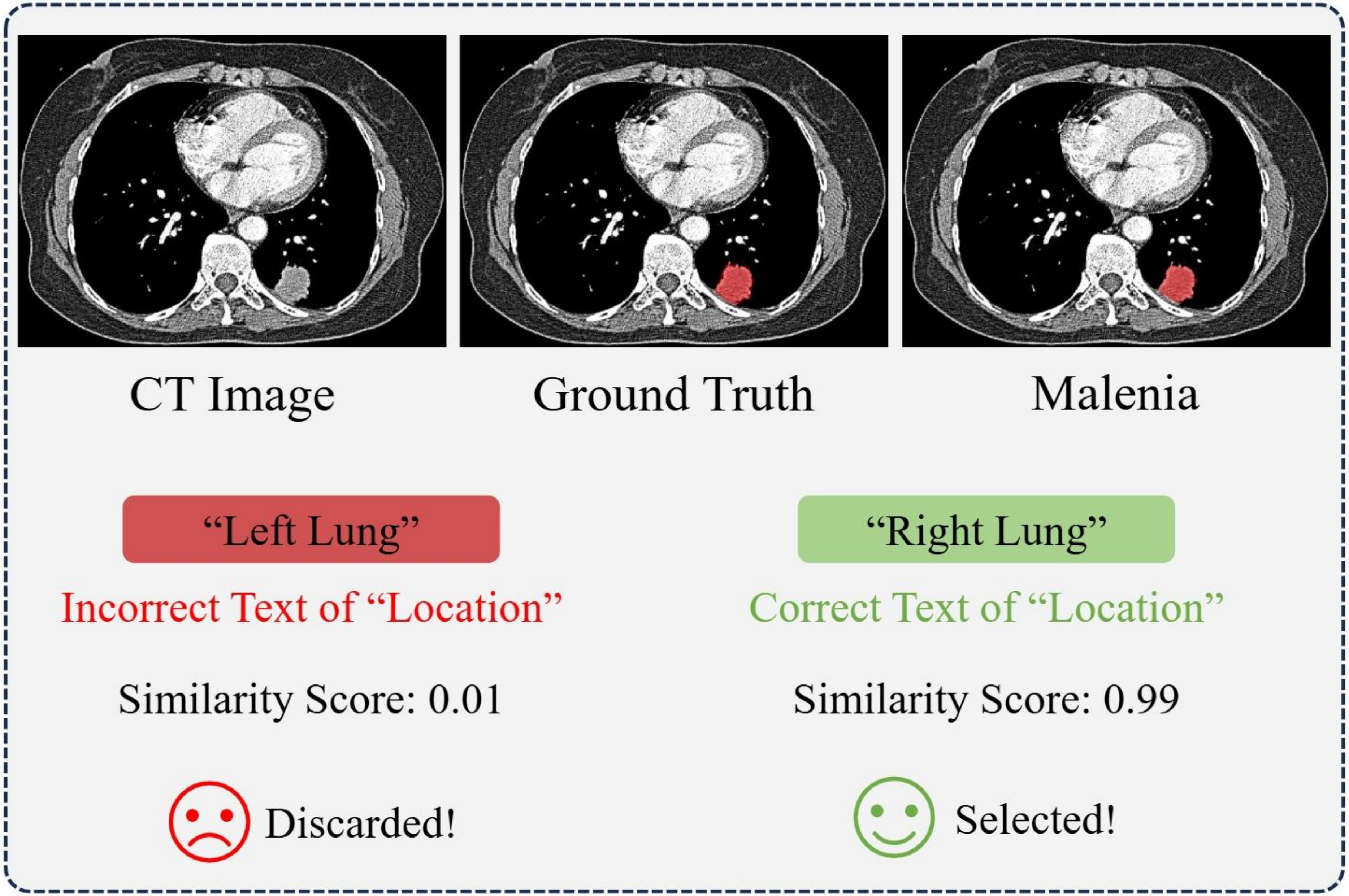




Results


CT Image
Ground Truth
Malenia
Location: "Liver" 1
Shape: "Round-like" 0.84
Density: "Hypodense Lesion" 0.87
Density Variation: "Heterogeneous" 0.80
Enhancement Status: "Enhanced CT" 1
Surface Characteristics: "Ill-Defined Margin" 0.77
Relationship with Surrounding Organs: "No Close Relationship with Adjacent Organs" 0.92
Specific Features: "Presence of Decreased Density Areas" 0.85
CT Image
Ground Truth
Malenia
Location: "Pancreas" 1
Shape: "Round-like" 0.94
Density: "Hypodense Lesion" 0.92
Density Variation: "Homogeneous" 0.90
Enhancement Status: "Enhanced CT" 1
Surface Characteristics: "Well-Defined Margin" 0.84
Relationship with Surrounding Organs: "Close Relationship with Adjacent Organs" 0.95
Specific Features: "Presence of Decreased Density Areas" 0.87
CT Image
Ground Truth
Malenia
Location: "Colon" 1
Shape: "Irregular Wall Thickening" 0.91
Density: "Hyperdense Lesion" 0.80
Density Variation: "Heterogeneous" 0.73
Enhancement Status: "Enhanced CT" 1
Surface Characteristics: "Ill-Defined Margin" 0.76
Relationship with Surrounding Organs: "Close Relationship with Adjacent Organs" 0.86
Specific Features: "Presence of Increased Density Areas" 0.83
CT Image
Ground Truth
Malenia
Location: "Right Lung" 1
Shape: "Irregular" 0.88
Density: "Low-Density Ground-Glass Opacity" 0.85
Density Variation: "Heterogeneous" 0.86
Enhancement Status: "Non-Contrast CT" 1
Surface Characteristics: "Ill-Defined Margin" 0.82
Relationship with Surrounding Organs: "No Close Relationship with Adjacent Organs" 0.97
Specific Features: "Spiculated Margins" 0.66

Results



CT Image
Ground Truth
Malenia
"Left Lung"
Incorrect Text of "Location"
Similarity Score: 0.01
Discarded!
"Right Lung"
Correct Text of "Location"
Similarity Score: 0.99
Selected!

Conclusion

Malenia is a novel vision-language pre-training method designed for 3D zero-shot lesion segmentation.

Future Work

- Cross-domain and cross-modality challenges
- Scalability to other imaging modalities
- Handling lesions with ambiguous boundaries
- Capturing complex visual semantics
- Generalization to diverse lesion types
- Catching up the performance of fully supervised models.

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Thanks!