AAPM 2024
66TH ANNUAL MEETING & EXHIBITION
JULY 21–25 | LOS ANGELES, CA
EMBRACING CHANGE.
IMPACTING PATIENT CARE.

Enhancing Swin Transformer With Semantic Attention For Explainable Prediction

Aneesh Rangnekar, Jue Jiang, Harini Veeraraghavan

July 21, 2024

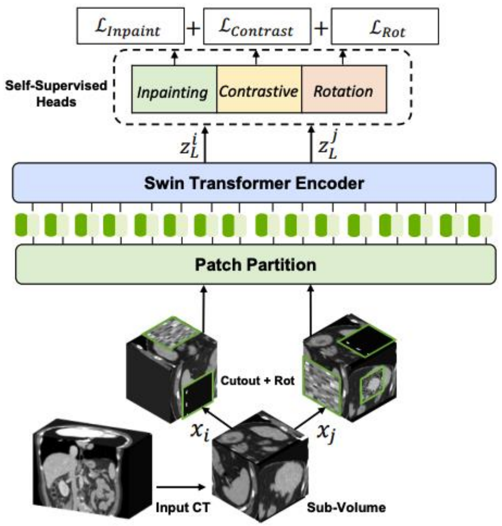
Session Title: Task-Based AI

Memorial Sloan Kettering
Cancer Center

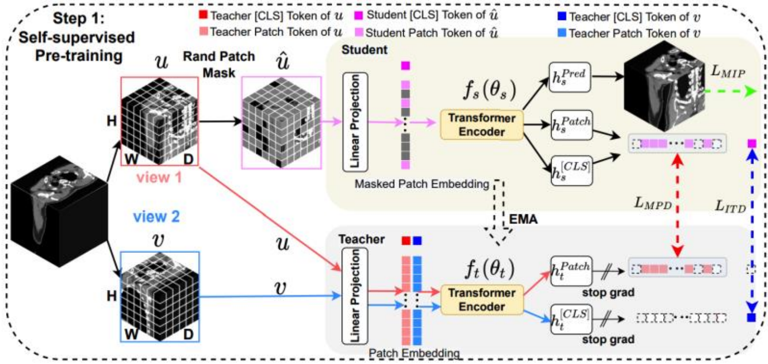
Motivation

- State-of-the-art foundation models in for 3D medical imaging
  - Swin UNETR
  - SMIT

  are based on the Swin transformer backbone
- However, the underlying architecture
  - cannot directly visualize the specific areas of interest
  - within the image scans
- We facilitate **visualization** of the regions of interest by modifying the architecture, thereby enabling eXplainable AI



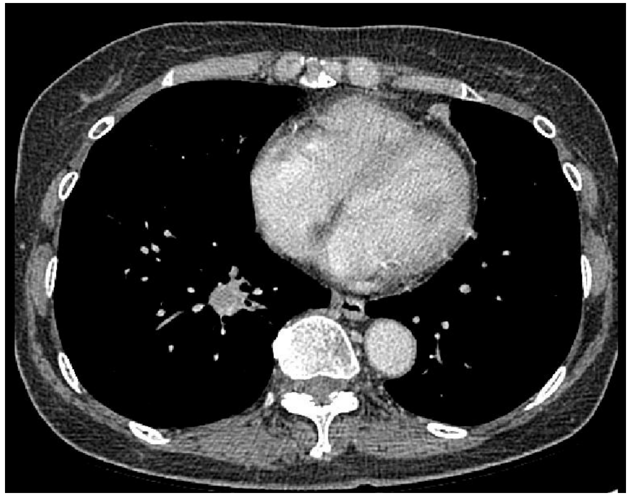
Swin UNETR



SMIT

Tang, Yucheng, et al. "Self-supervised pre-training of swin transformers for 3d medical image analysis." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
Jiang, Jue, et al. "Self-supervised 3D anatomy segmentation using self-distilled masked image transformer (SMIT)." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2022.

How do [Vision] transformers work?

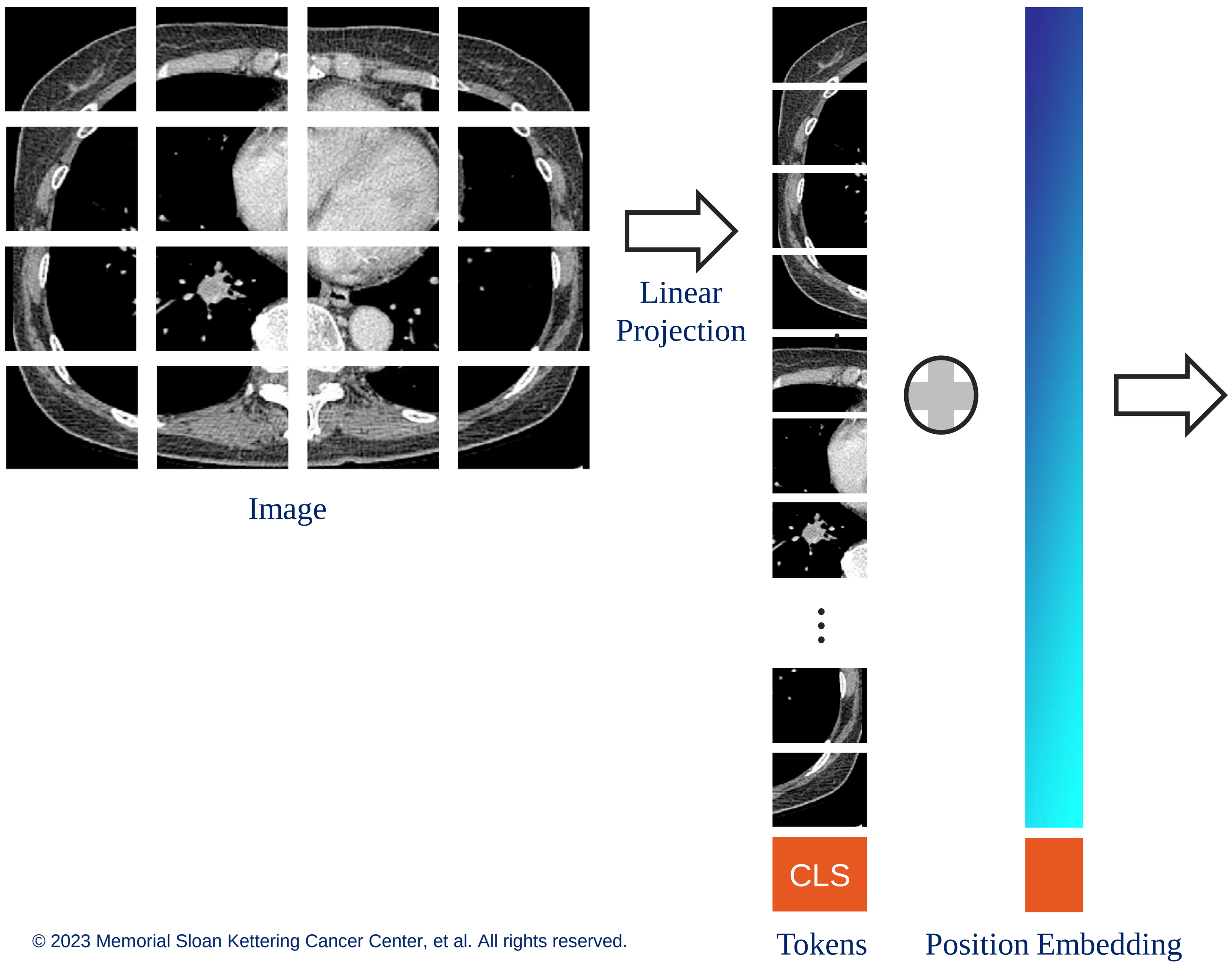


Image*

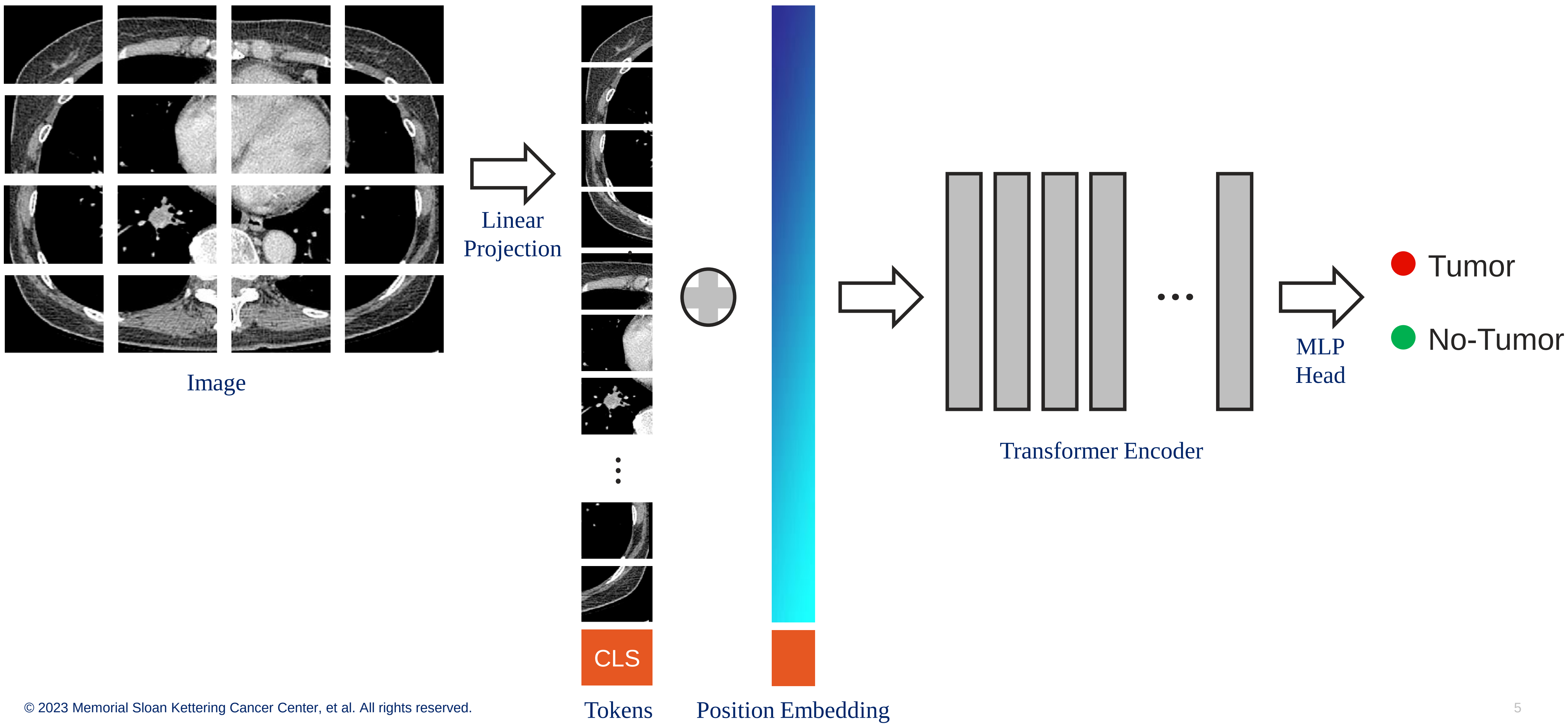
*Slice from 3D Volume for simplicity

Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." Proceedings of the International Conference on Learning Representations (ICLR), 2021.

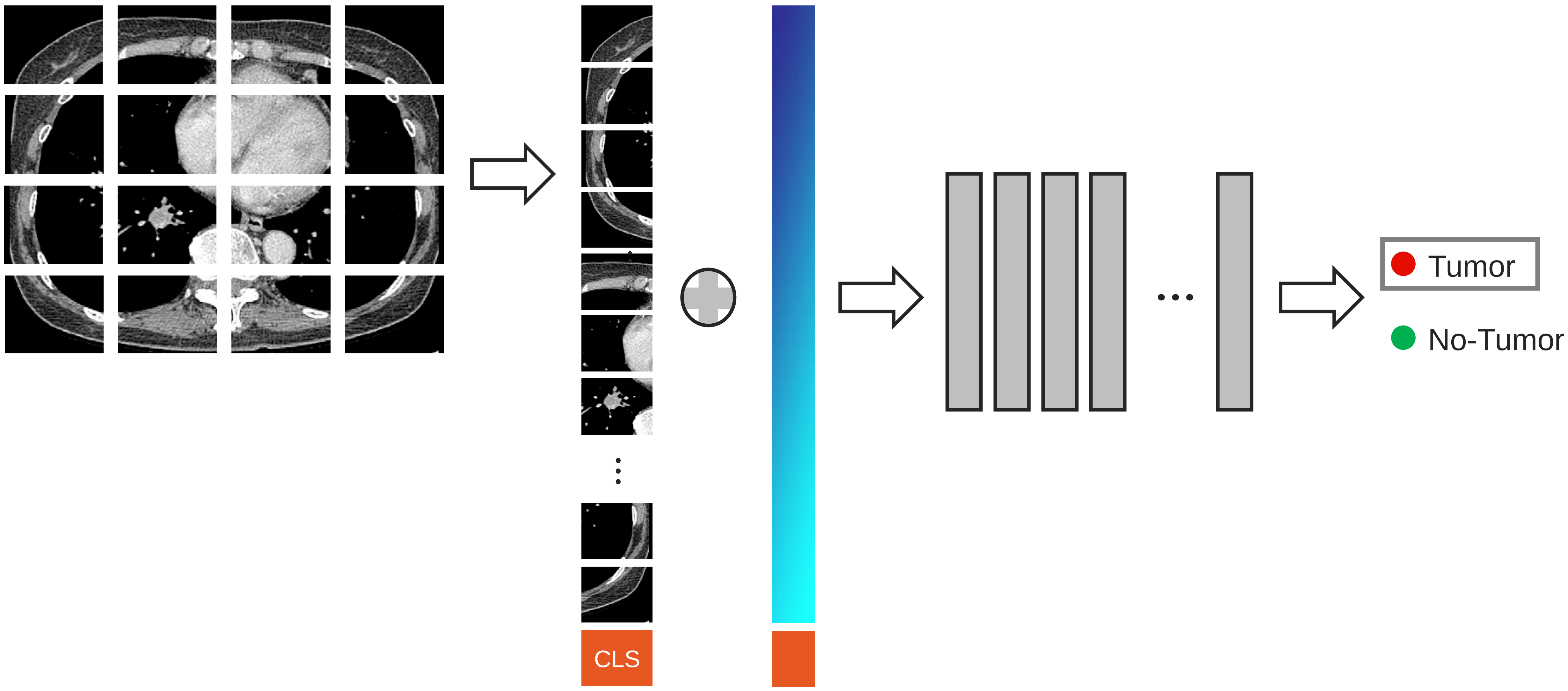
How do [Vision] transformers work?



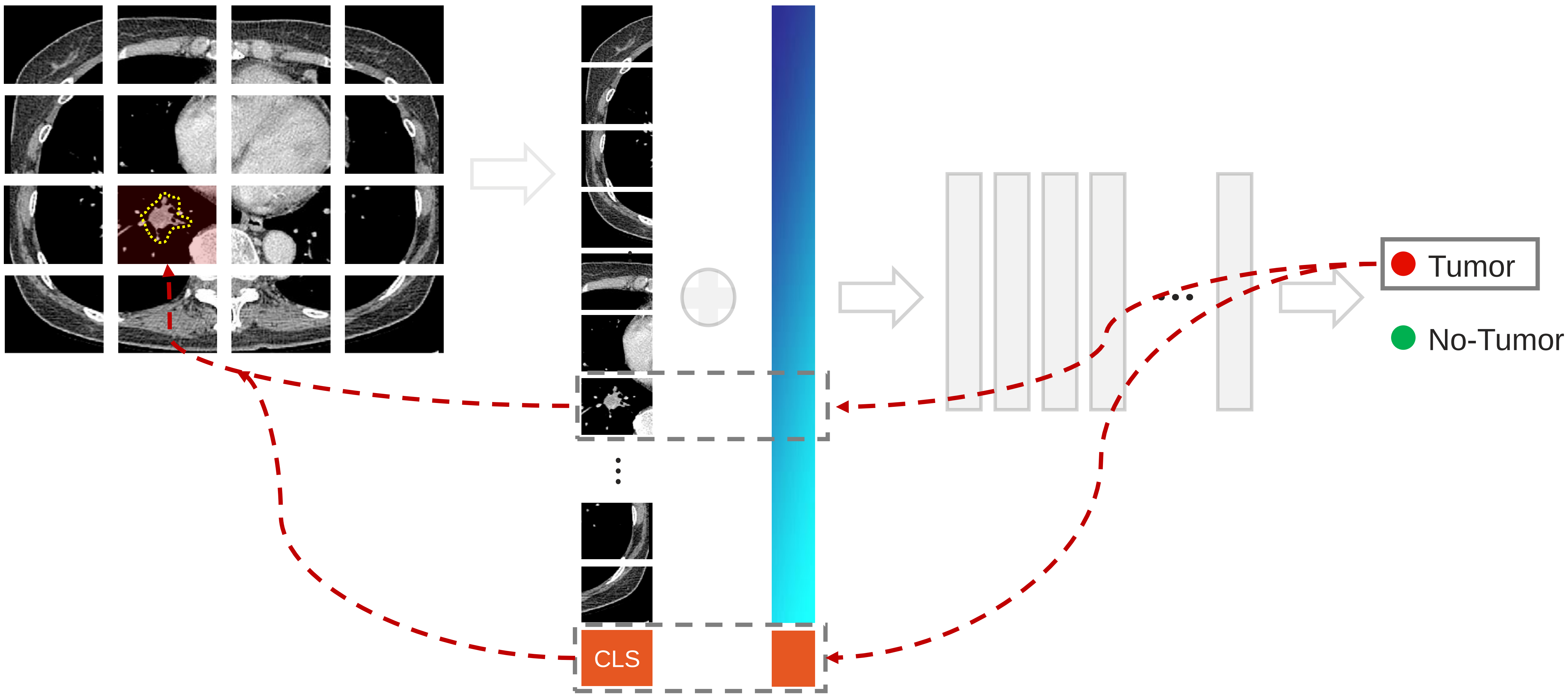
How do [Vision] transformers work?



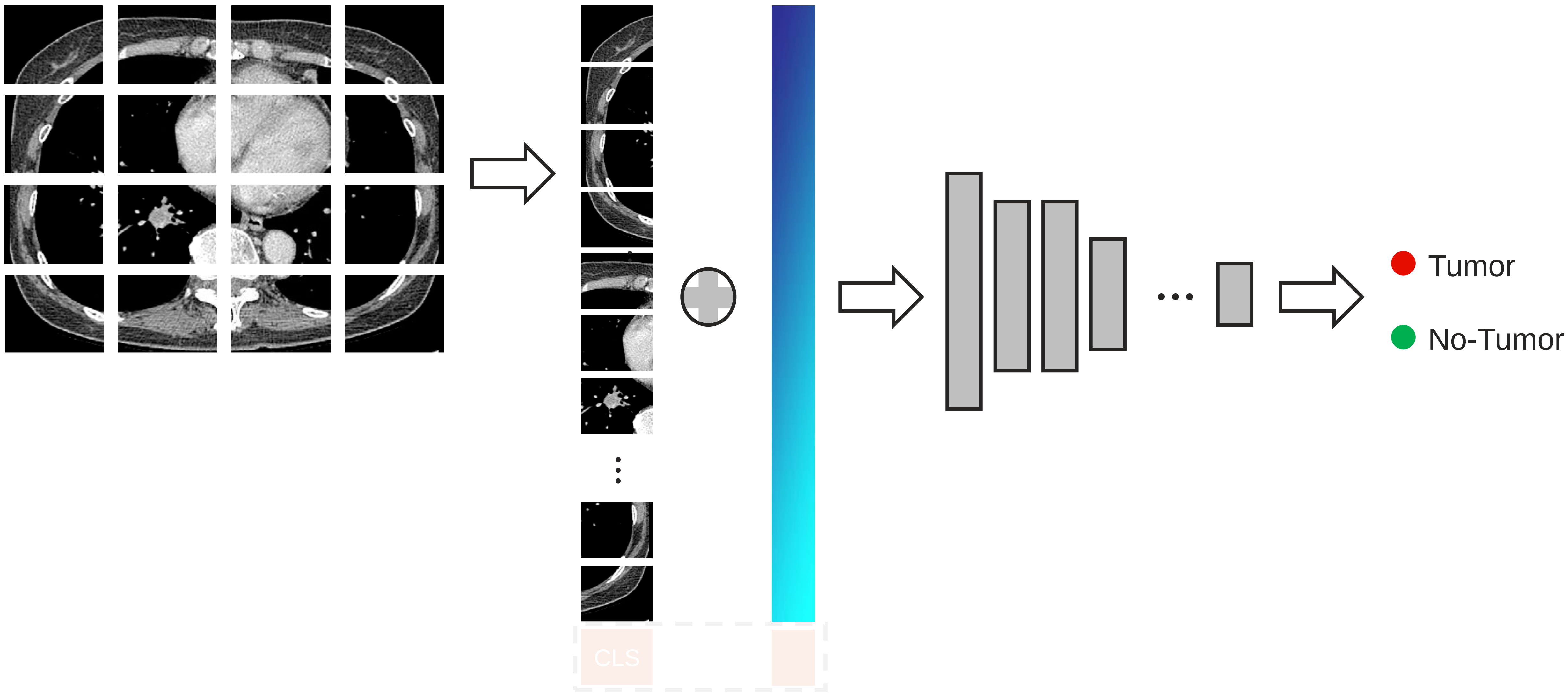
How do [Vision] transformers work?



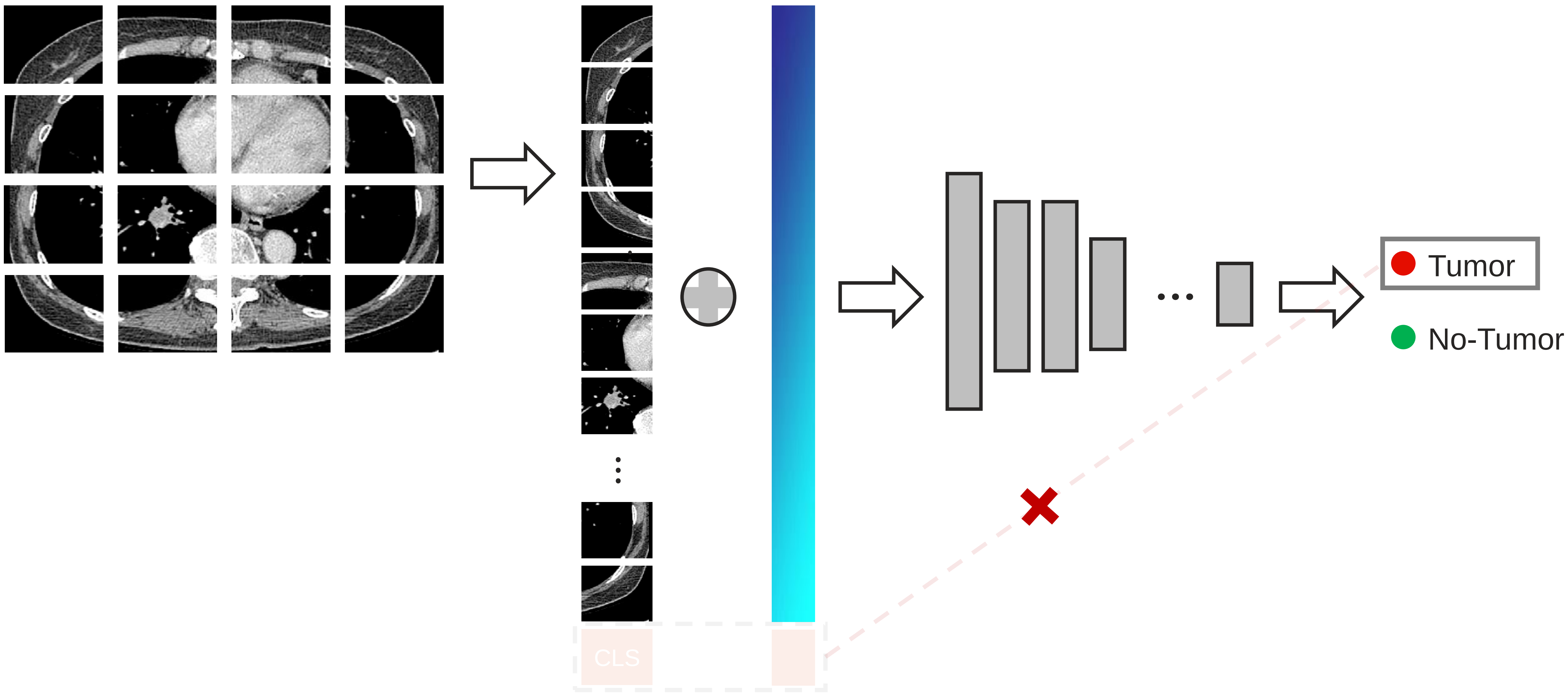
How do [Vision] transformers work?



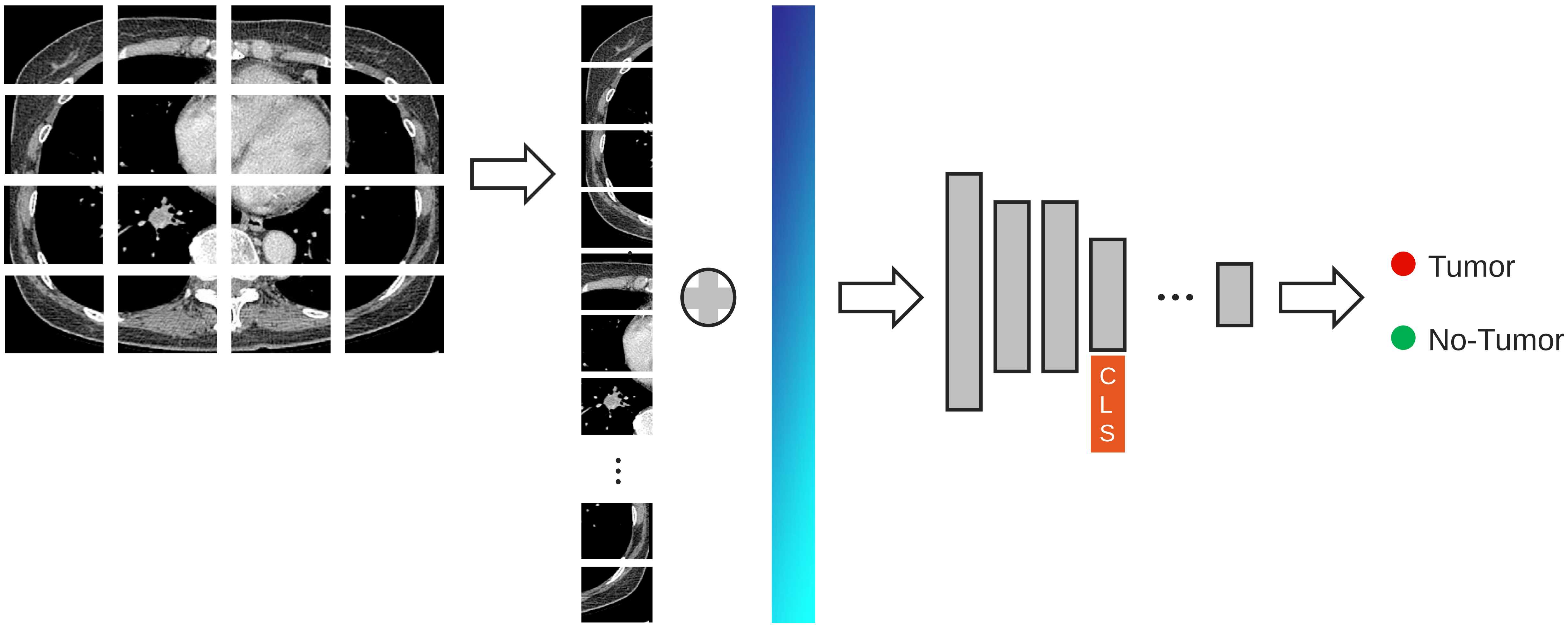
How do [Swin] transformers work?



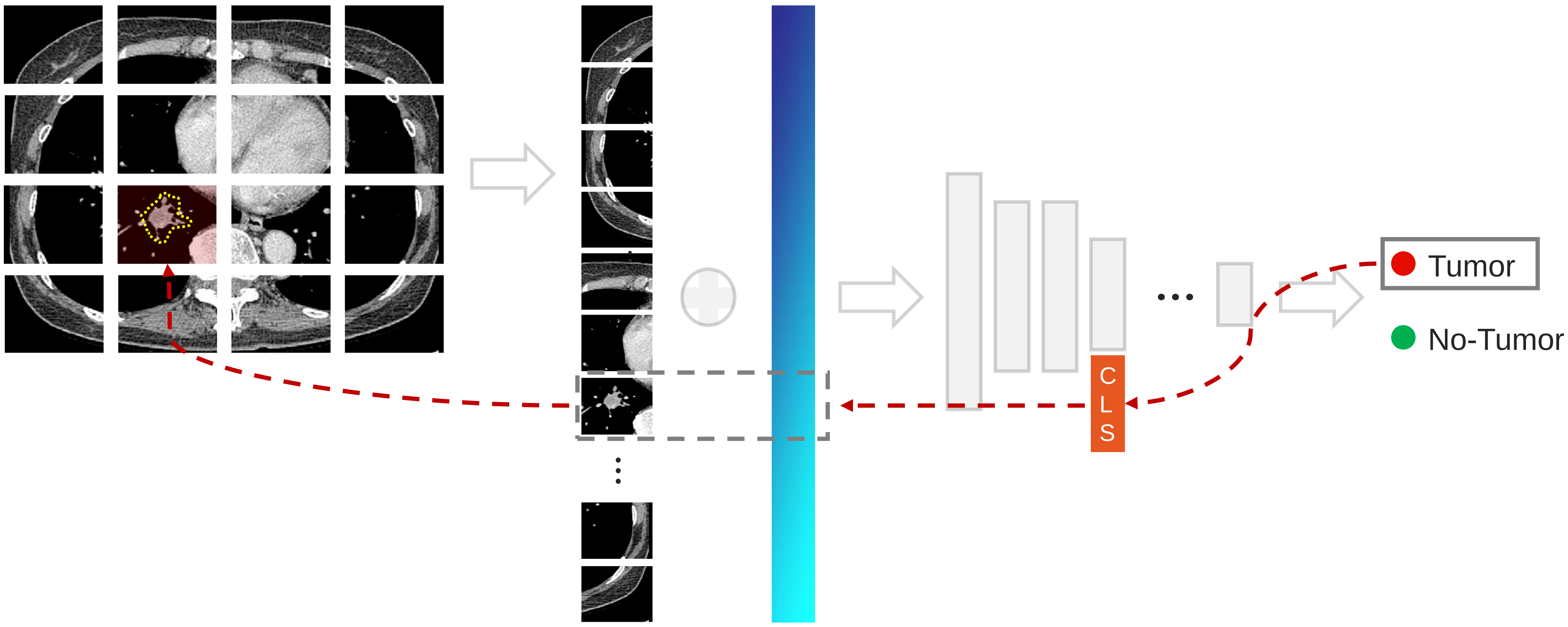
How do [Swin] transformers work?



How do we enhance the [Swin] transformer?

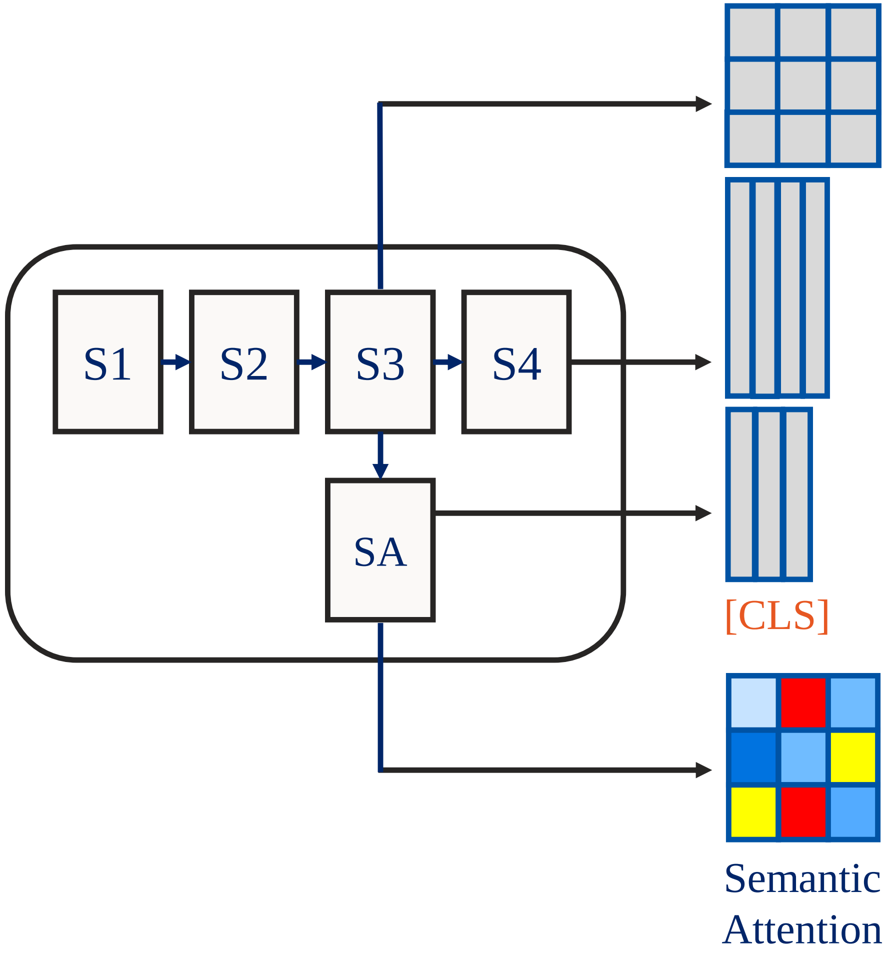


How do we enhance the [Swin] transformer?



Foundation Model: SMART

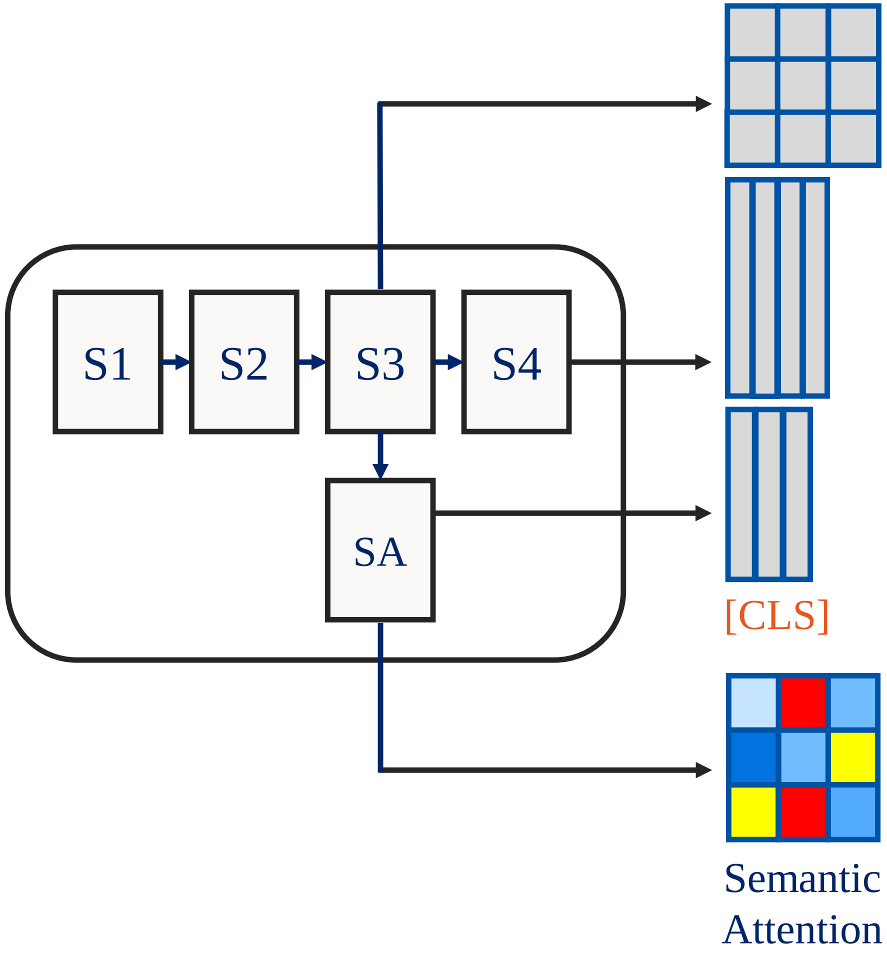
[S]elf-distilled [M]asked [A]ttention guided masked image modeling with noise [R]egularized [T]eacher



*built on SMIT

Jiang, Jue, et al. "Self-supervised 3D anatomy segmentation using self-distilled masked image transformer (SMIT)." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2022.

Foundation Model: SMART



Student → Noisy Teacher

10k 3D CT Scans:
- Lung Cancer
- Head and Neck
- Multi-organ Abdominal
- Kidney
- Liver
- Pancreas
- COVID-19

Step 1: Pretraining with Masked Image Modeling and Token Distillation

Student → Noisy Teacher

Immunotherapy Dataset
- Durable Clinical Benefit Prediction
- Tumor Segmentation

Step 2: Fine-Tuning

He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

Results

We compare SMART against two other approaches:

- AttMask
- SMIT

Pretext Task	Linear Probing			Fine-tuning		
	AP_{50}	AR_{50}	AUC	AP_{50}	AR_{50}	AUC
AttMask	44.9	54.0	0.570	56.0	68.5	0.660
SMIT	46.4	58.1	0.620	56.3	67.0	0.660
SMART	54.5	68.4	0.660	57.4	71.7	0.740

Durable Clinical Benefit Prediction with 100% data utilization

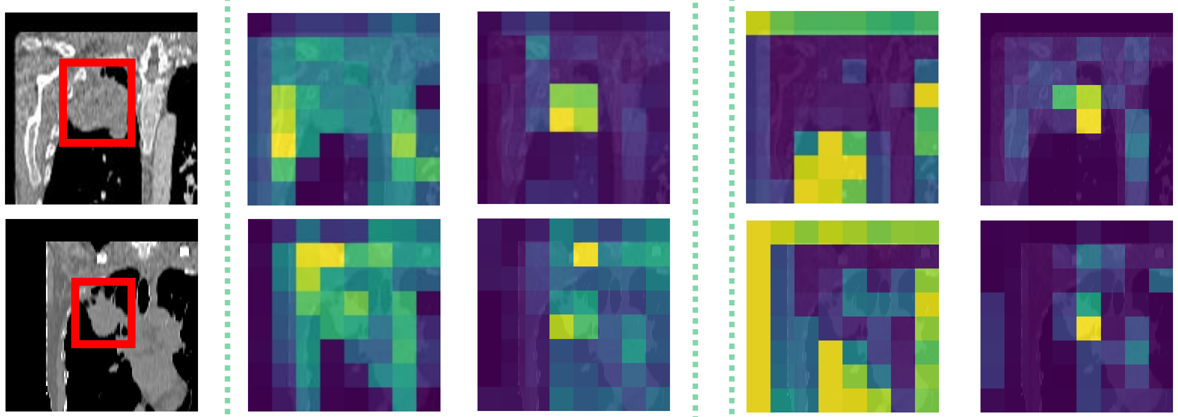
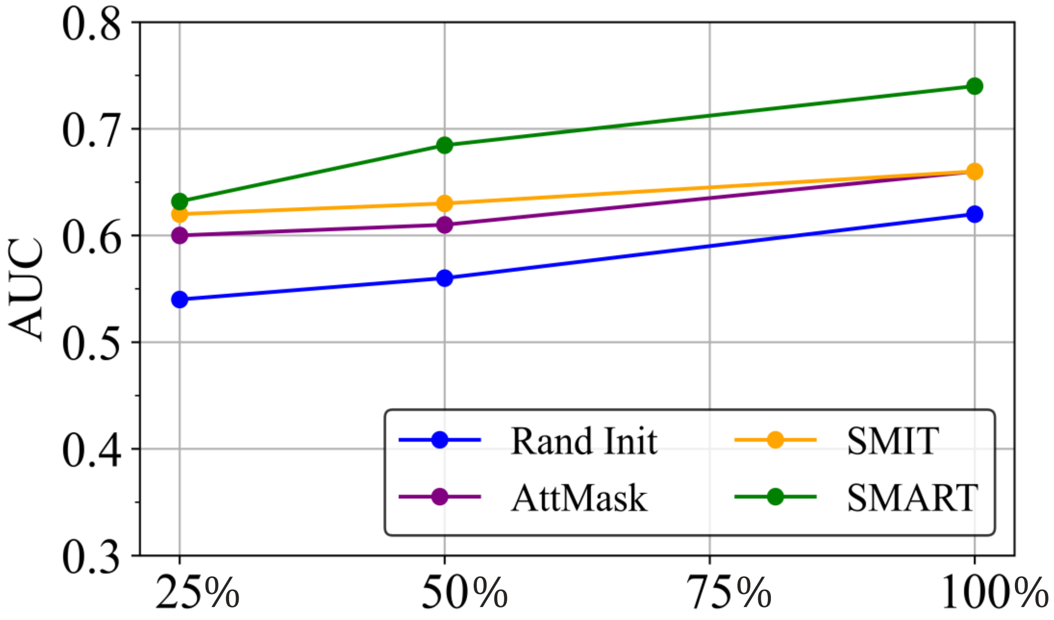


Image AttMask SMART AttMask SMART

Prediction Attention Visualization post Fine-Tuning

Kakogeorgiou, Ioannis, et al. "What to hide from your students: Attention-guided masked image modeling." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.
Jiang, Jue, et al. "Self-supervised 3D anatomy segmentation using self-distilled masked image transformer (SMIT)." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2022.

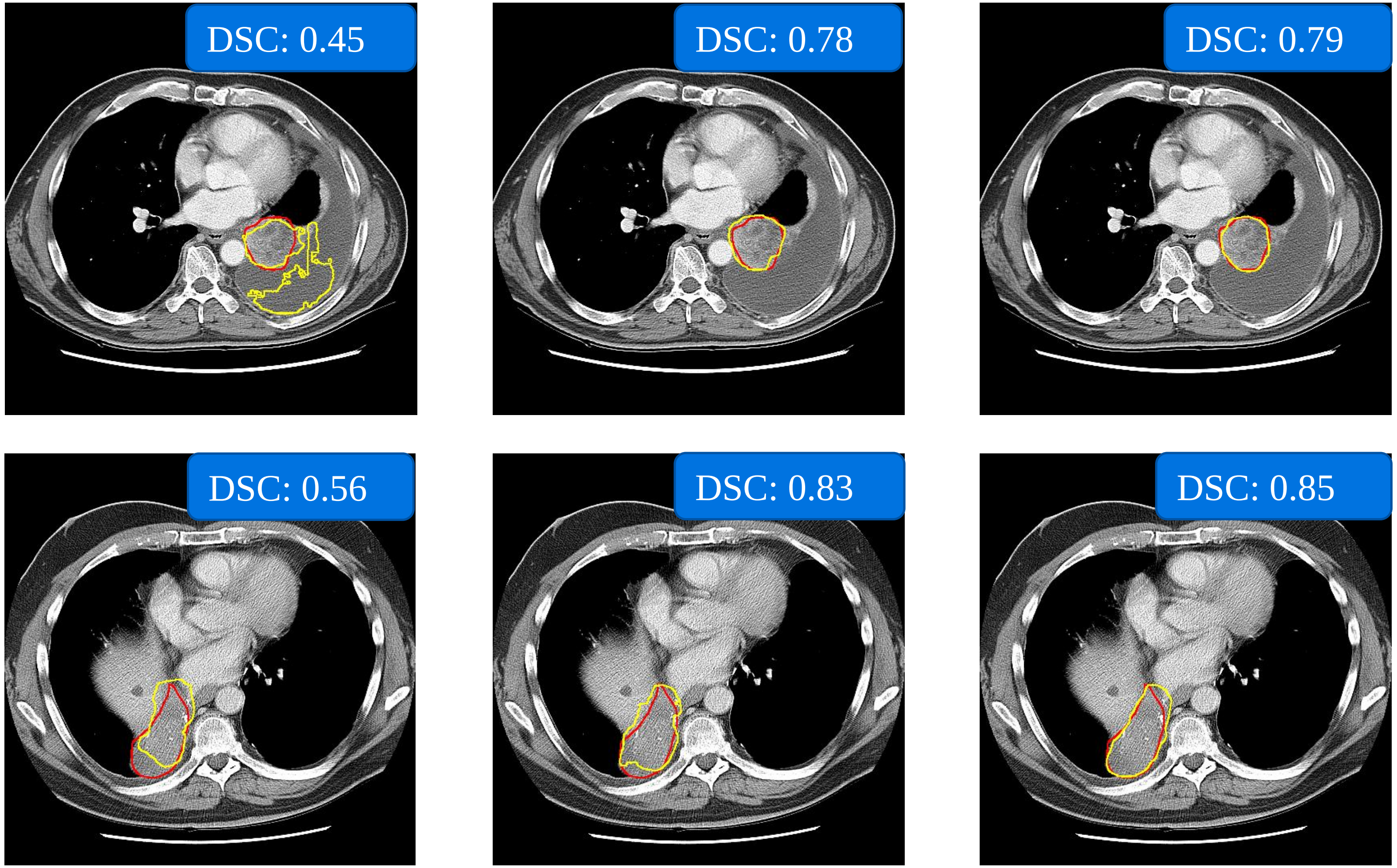
Results



Durable Clinical Benefit Prediction with limited-data learning protocol

Results

Tumor Segmentation	
AttMask	0.69 ± 0.21
SMIT	0.76 ± 0.13
SMART	0.77 ± 0.11

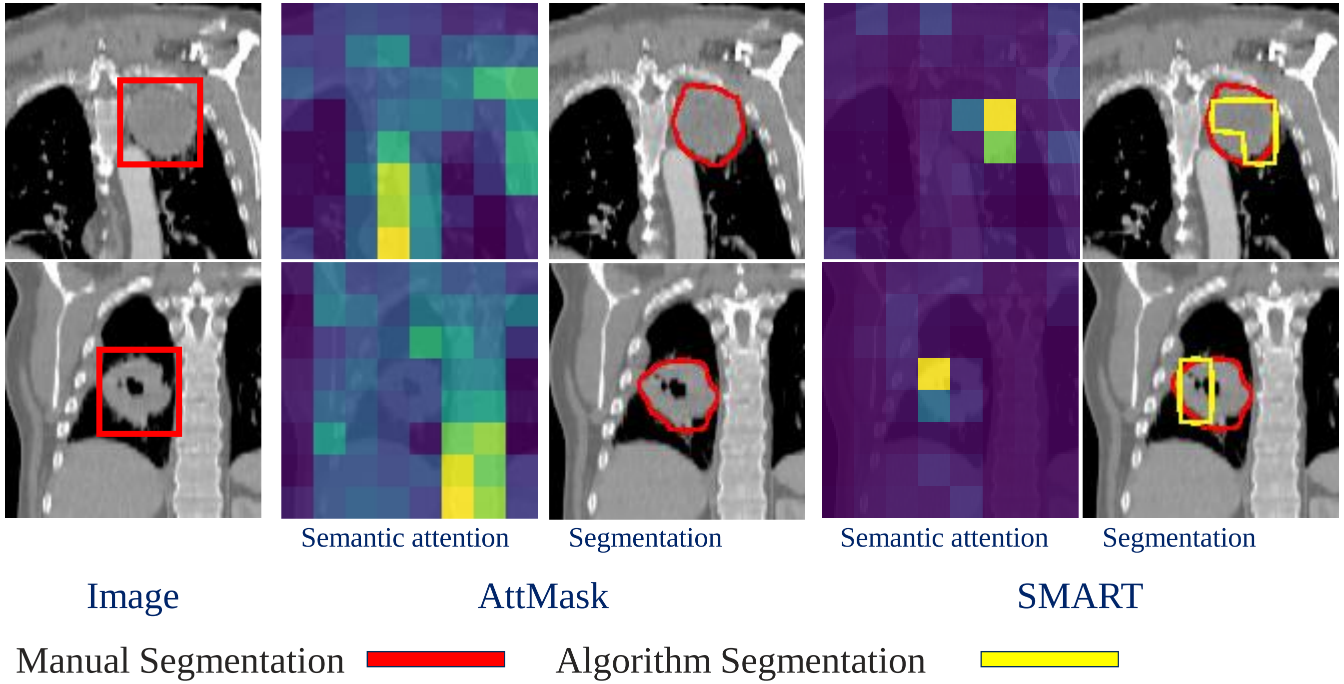


AttMask SMIT SMART

Manual Segmentation Algorithm Segmentation

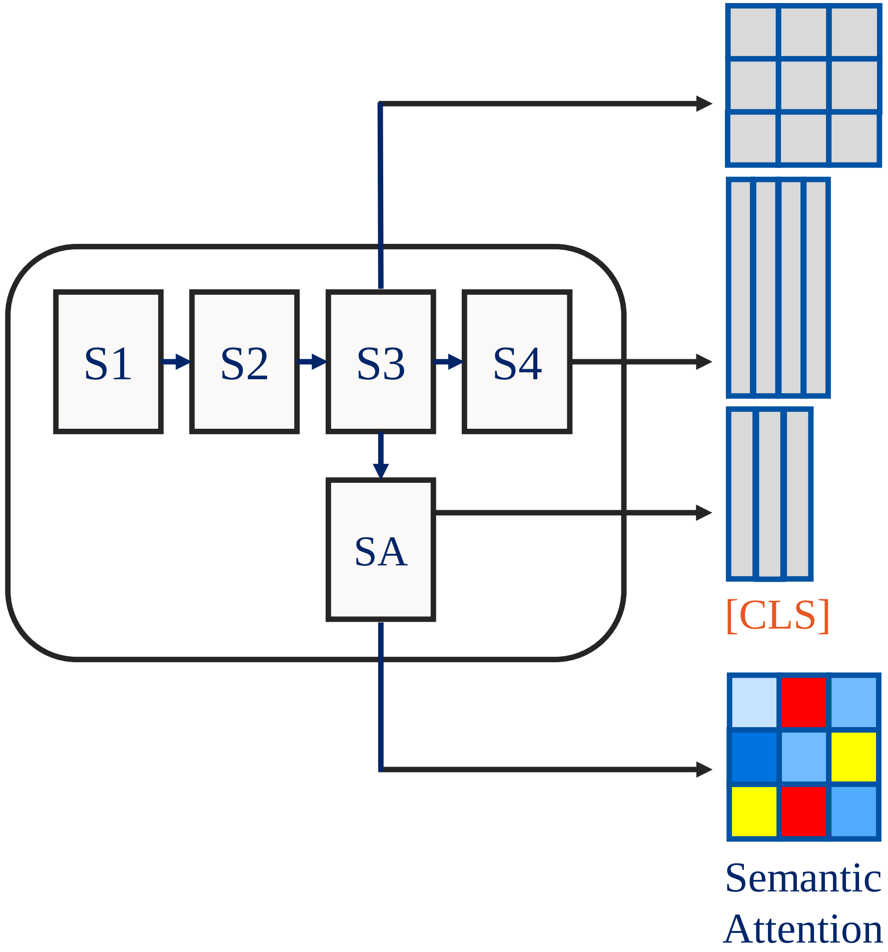
Results

Additional Results for zero-shot localization on held-out TCIA dataset



Summary

- We introduced a semantic attention block that enables CLS token
  - Helps pretraining the network with informative token masking
  - Helps post fine-tuning XAI
- Further boosts fine-tuning performance under limited data
- Also applied to
  - LIDC dataset for tumor malignancy classification and segmentation
  - TCIA dataset for zero-shot tumor localization
  - Lung Radiomics and Radiogenomics dataset for tumor segmentation
  - on arXiv: 2310.01209



Thank you

Questions?