Kidney Cancer Detection from CT Images by Transformer-Based Classifiers

Toru Tanaka[*1] Takaki Fukazawa[*1] ○Yoshitaka Kameya[*1]
Keiichi Yamada[*1] Kazuhiro Hotta[*1] Tomoichi Takahashi[*2]
Naoto Sassa[*3] Yoshihisa Matsukawa[*4]
Shingo Iwano[*4] Tokunori Yamamoto[*4]

[*1] Meijo University
[*2] Meis Technology Inc.
[*3] Aichi Medical University School of Medicine
[*4] Nagoya University Graduate School of Medicine

Outline

- Background
- Methods
- Experimental Results
- Conclusion

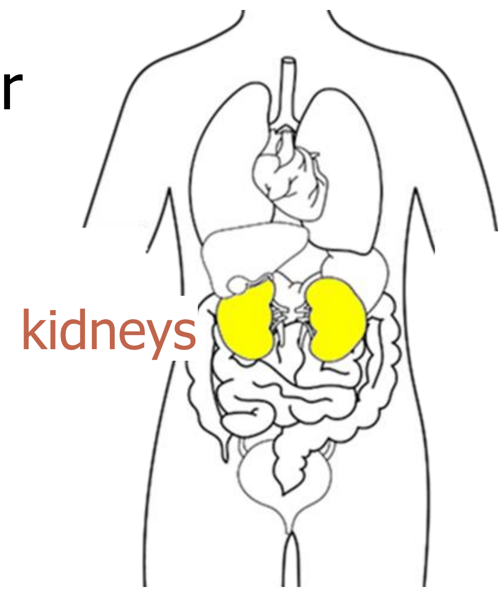
Outline

- Background
- Methods
- Experimental Results
- Conclusion

Background: Kidney Cancer Detection (1)

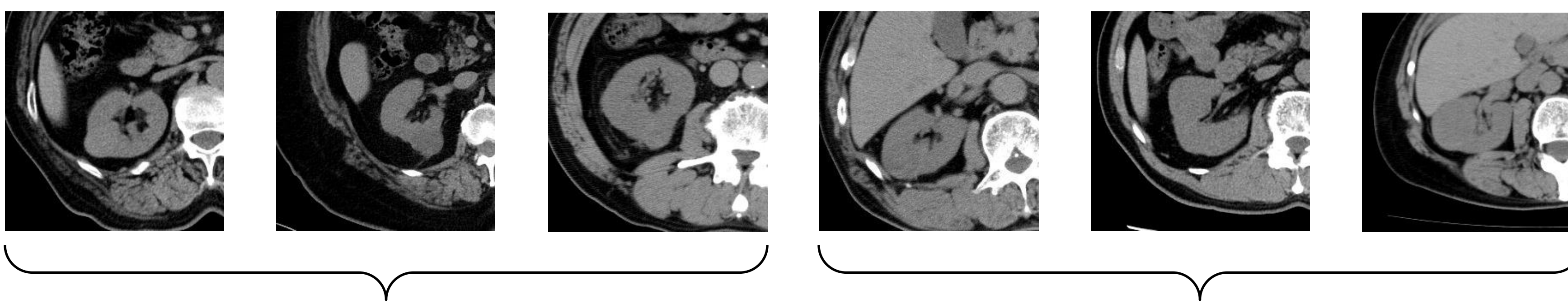
- Convolutional neural networks (CNNs) have been widely adopted in medical image analysis
- CNNs have also been applied to diagnoses on kidney cancer based on abdominal CT images:
  - Detecting kidney cancer [Hussain+ 17][Takahashi+ 20]

    Our target
  - Discriminating between benign and malignant renal masses [Oberai+ 20]
  - Identifying the subtypes of kidney cancer [Han+ 19][Uhm+ 21]



Background: Kidney Cancer Detection (2)

- Challenges:
  - Texture of cancerous tumors and normal tissues can be very similar
  - Locations of abdominal organs can vary depending on individual patients
  - Some cancerous tumors (called endophytic tumors) can grow inside kidneys



Exophytic tumors | Endophytic tumors

Background: Kidney Cancer Detection (3)

- Human experts conventionally use some substances called contrast agents to enhance the contrast among tissues

30 seconds after injecting a contrast agent to the patient

UCT (Unenhanced CT) image

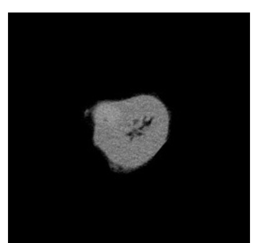
CECT (Contrast-enhanced CT) image

- Some patients have allergy to contrast agents
- Contrast agents may worsen the renal function

High clinical cost

- It was reported that masking all organs around a kidney is effective [Takahashi+ 20]

Mask



High annotation cost

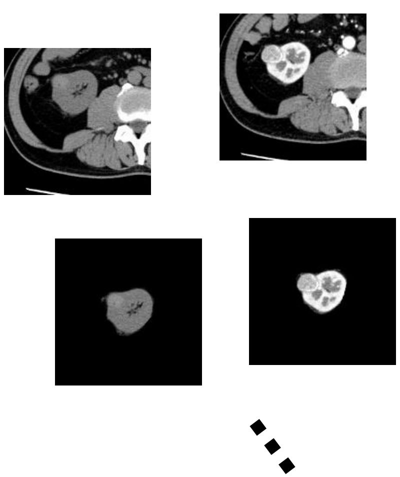
Background: Our Motivation

- In NLP and CV, Transformer-based deep neural networks have demonstrated high predictive performance

This work:

We apply transformer-based classifiers to kidney cancer detection from various types of CT images

Various types of CT images with different clinical/annotation costs



Transformer-based classifiers

Vision Transformer (ViT)
Swin Transformer

Detection:

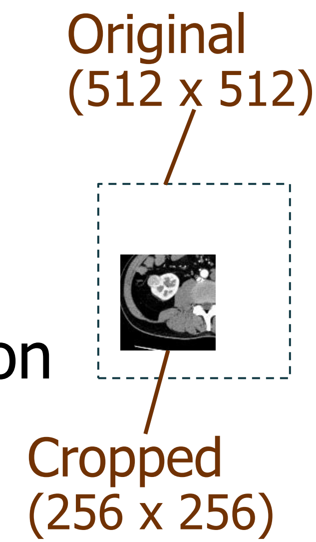
Presence/absense of a cancerous tumor

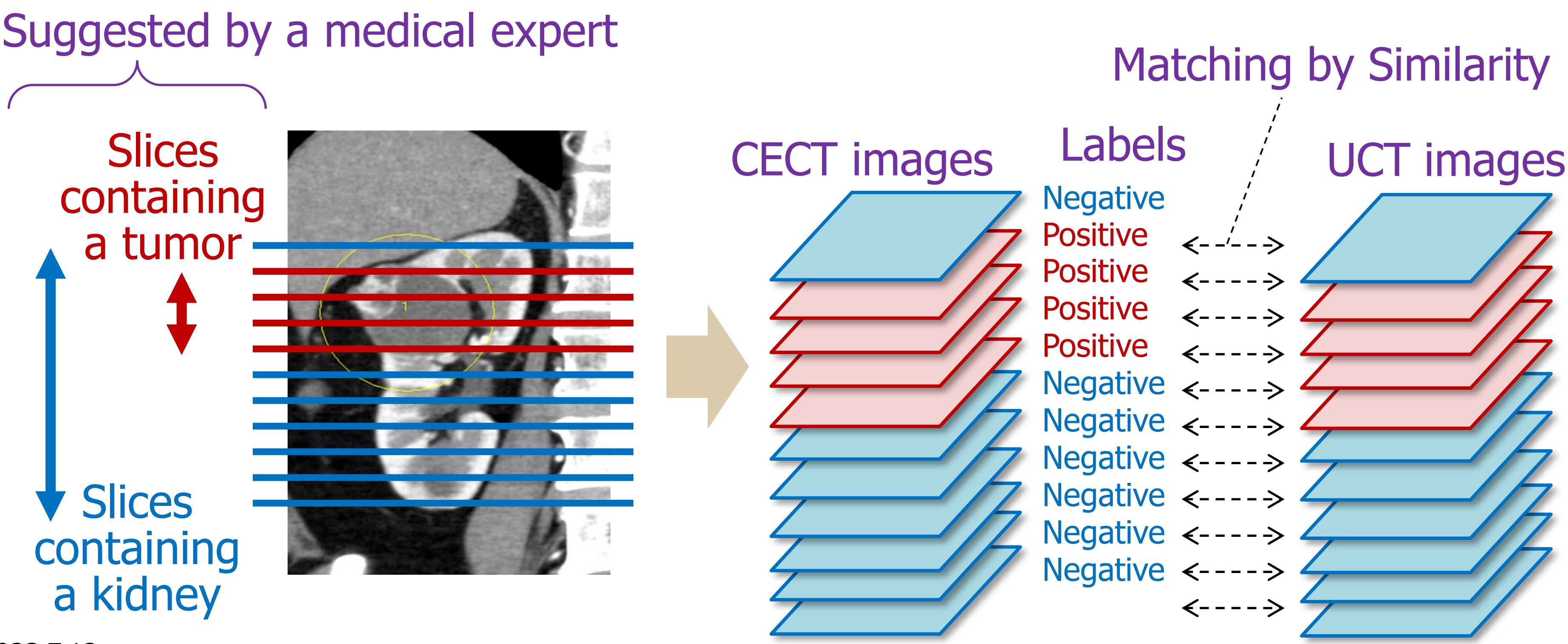
Outline

- ✓ Background
- Methods
- Experimental Results
- Conclusion

Methods: Datasets (1)

- CT values stored in DICOM-formatted files were converted into digital images
- Area of size 256 x 256 was cropped at a fixed position where the majority of kidneys were centered
- CT images were labelled as positive or negative:





Methods: Datasets (2)

- CT images containing a left kidney were horizontally flipped
  - Supposed that right and left kidneys are symmetric
  - One patient having two kidneys = Two virtual patients

- Entire dataset was split in our experiment:

Dataset	# of (virtual) patients		# of CT images	
	Present	Absent	Present	Absent
Training	148	146	728	5,212
Validation	30	30	141	1,027
Evaluation	39	40	196	1,397
Total	217	216	1,065	7,636

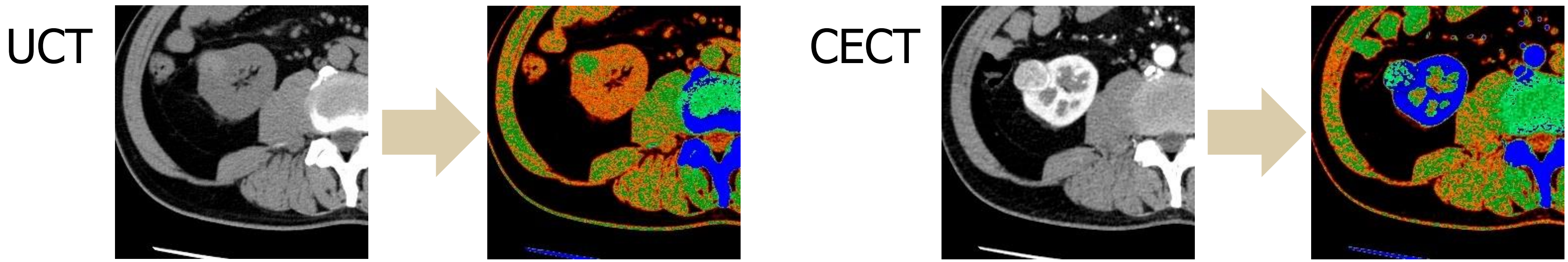
433

8,701

- Data augmentation at the beginning of each epoch
  - Shift, Rotation, Shear transformation, and Zooming-in/out

Methods: Datasets (3)

- Virtually colorized version of all CT images were created based on the CT values in the Hounsfield Unit scale

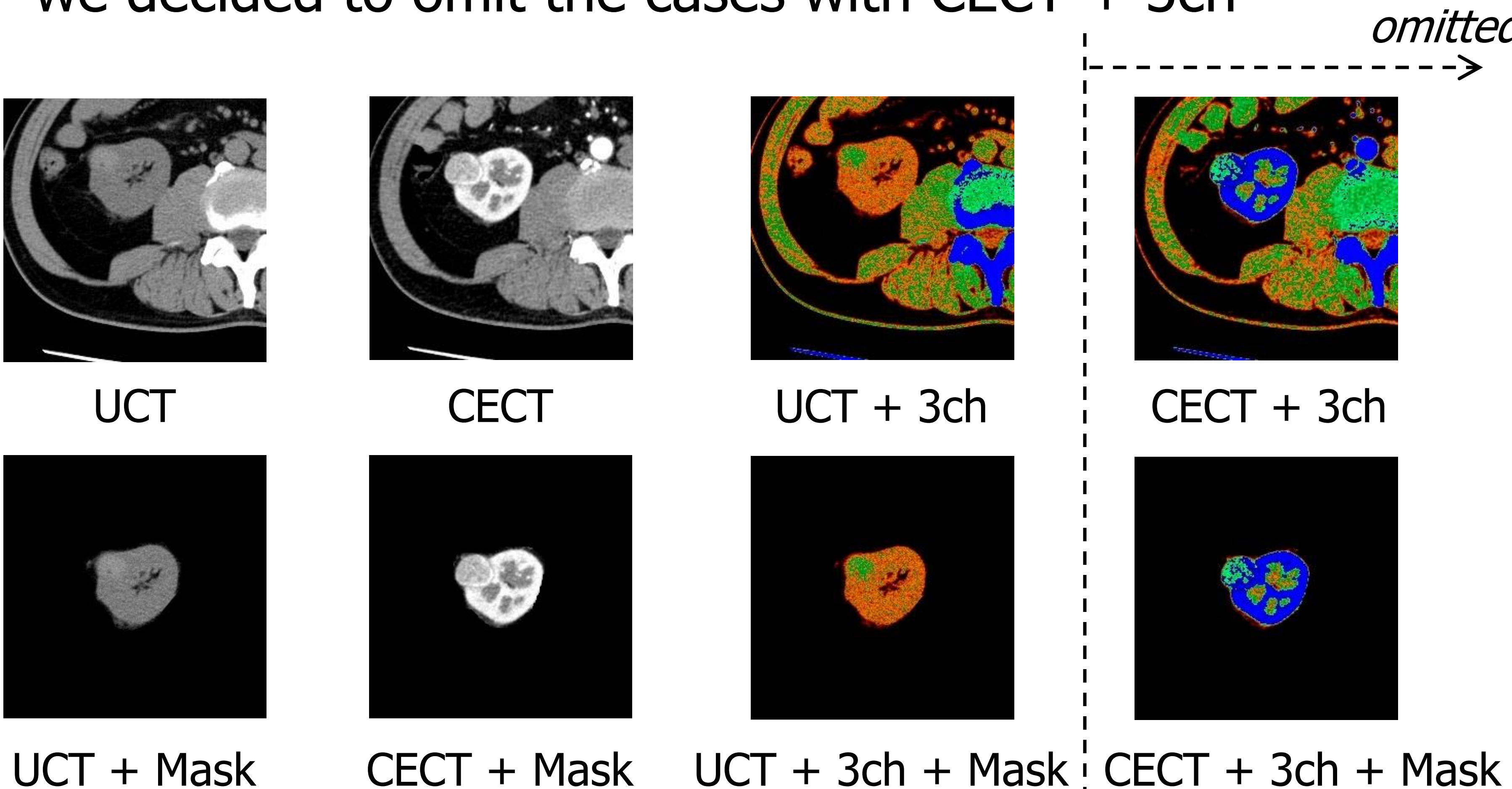


		Range of CT value (HU)	Tissue
Grayscale		−110 ~ 190	All
Virtual color (3ch) [Takeuchi+ 21]	Red	−70 ~ 50	Fat and water
	Green	−10 ~ 110	Water and soft tissue
	Blue	50 ~ 170	Soft tissue and bone

- Masked version of all CT images were created manually

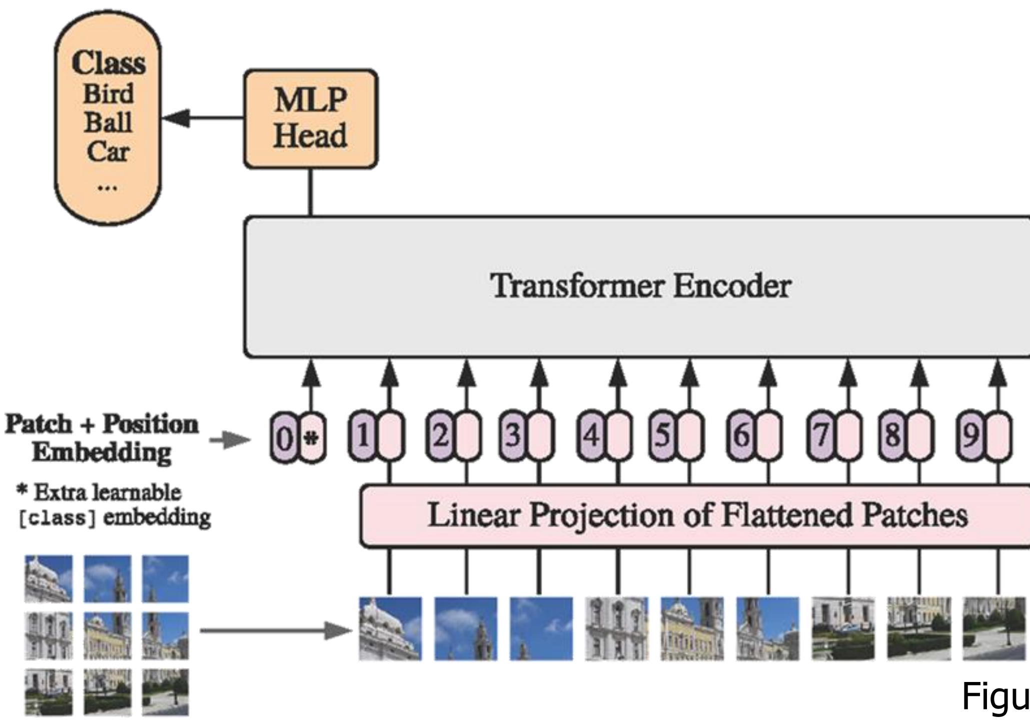
Methods: Datasets (4)

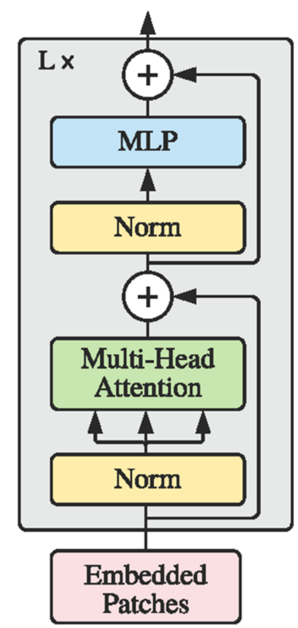
- Consequently, we can consider 8 variations of CT images with different clinical/annotation costs
- From the results of a preliminary experiment, we decided to omit the cases with CECT + 3ch



Methods: DNNs for Image Classification (2)

- Convolutional neural networks (CNNs)
  - We used VGG-16, ResNet-50
- Vision Transformer (ViT) [Dosovitskiy+ 21]
  - Uses the encoder part of the original Transformer
  - Splits an input image into patches of a fixed size
  - Applies multi-head self-attention (MSA) repeatedly
  - Can capture long-range dependency in the input image

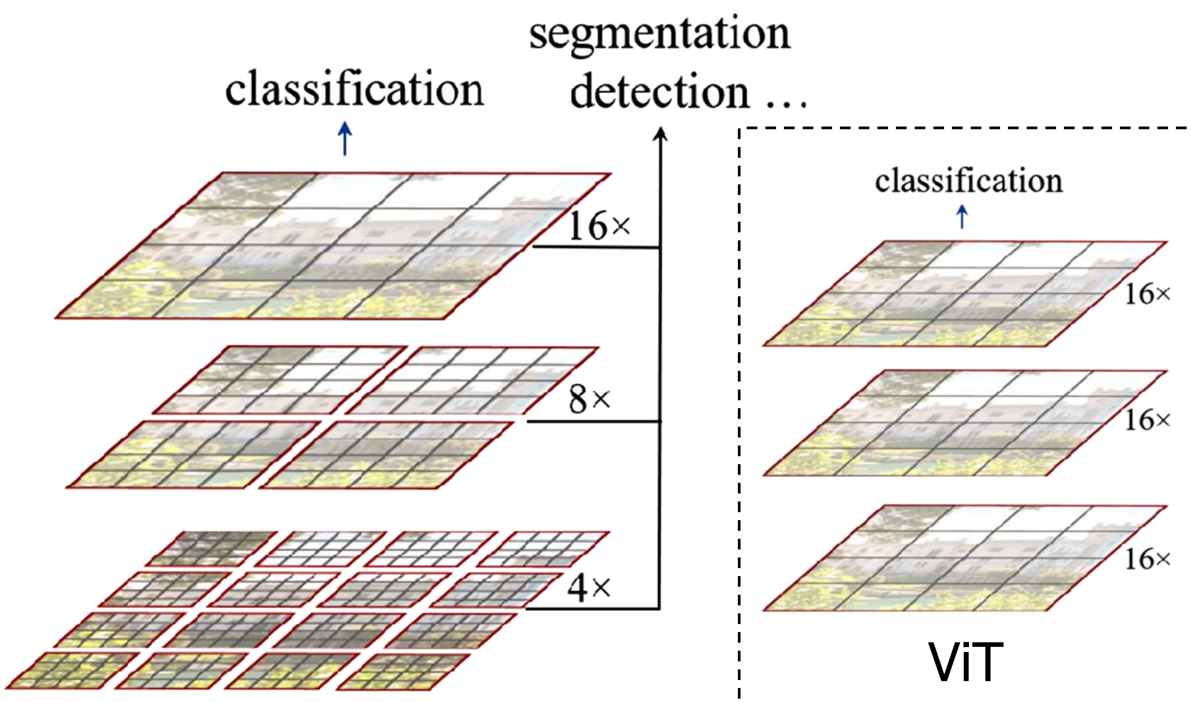




Figures excerpted from [Dosovitskiy+ 21]

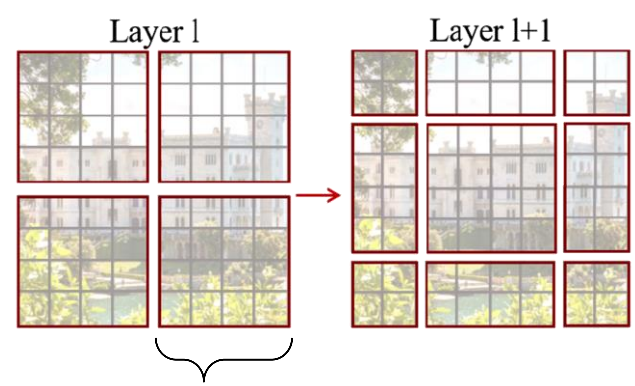
Methods: DNNs for Image Classification (3)

- Swin Transformer [Liu+ 21]
  - Inherits the basic architecture from ViT
  - Introduces a CNN-like hierarchical and local structure via patch merging
  - Performs window-based MSA (W-MSA) and shifted window-based MSA (SW-MSA) alternately



Swin Transformer

W-MSA SW-MSA

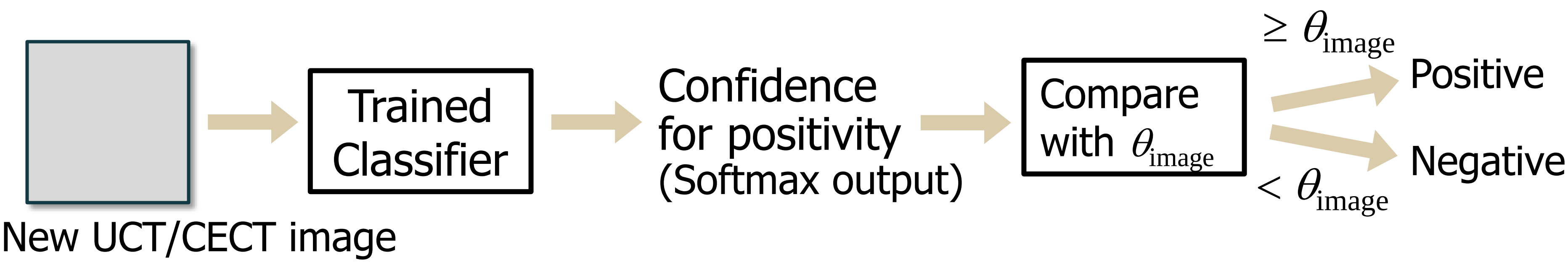


Window
= Range of a self-attention being performed

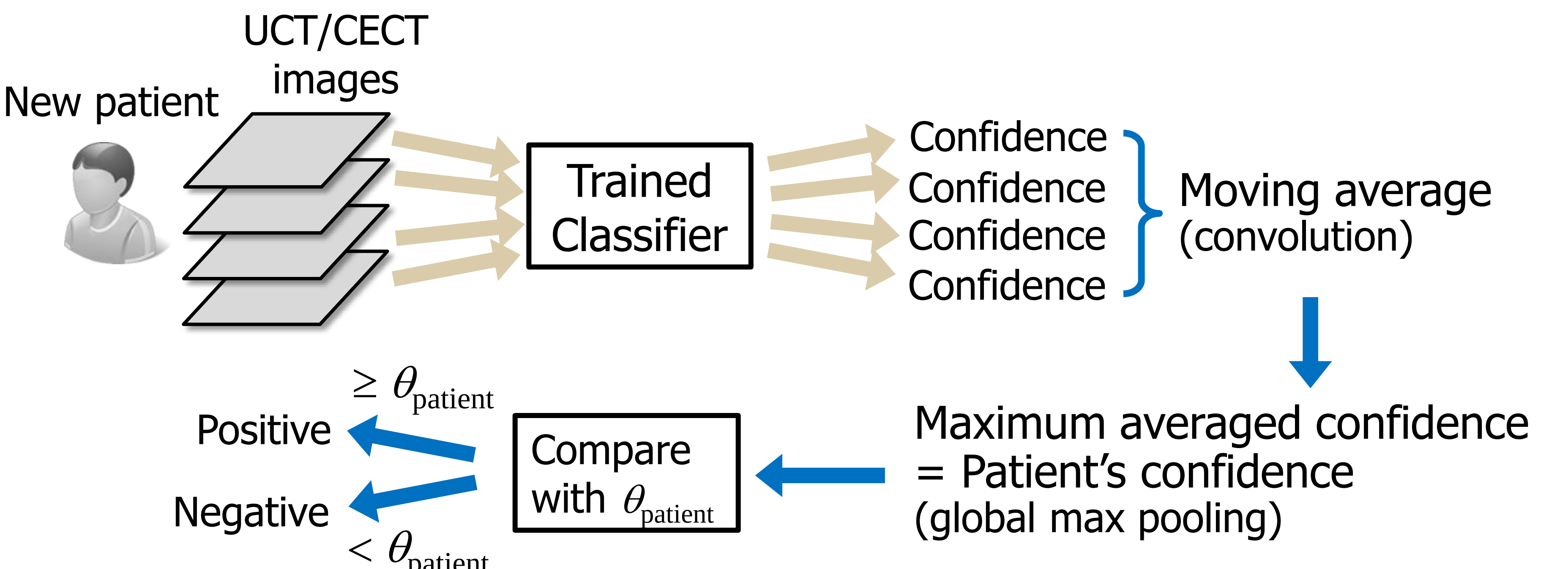
Figures excerpted from [Liu+ 21]

Methods: Image/Patient-wise Detection

- Image-wise detection



- Patient-wise detection



Methods: Configuration for Training

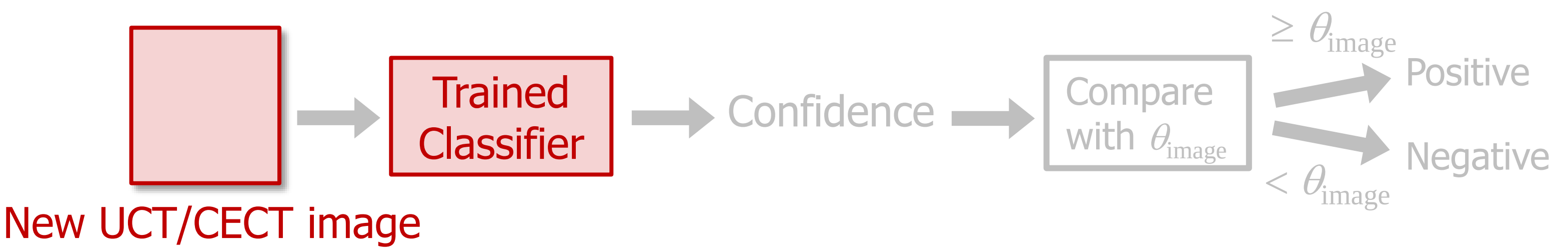
- Loss: Weighted cross-entropy (for coping with class imbalance)
- AdaGrad with initial learning rate of 10^{-5}
- Mini-batch size: 32
- # of epochs:
  - 150 for VGG-16/ResNet-50 with UCT images
  - 200 for VGG-16/ResNet-50 with CECT images
  - 50 for ViT
  - 100 for Swin Transformer
- Pre-trained models:
  - VGG-16 with ImageNet-1K (# of parameters: 134M)
  - ResNet-50 with ImageNet-1K (24M)
  - ViT with ImageNet-21K (303M)
  - Swin Transformer with ImageNet-21K (195M)

Outline

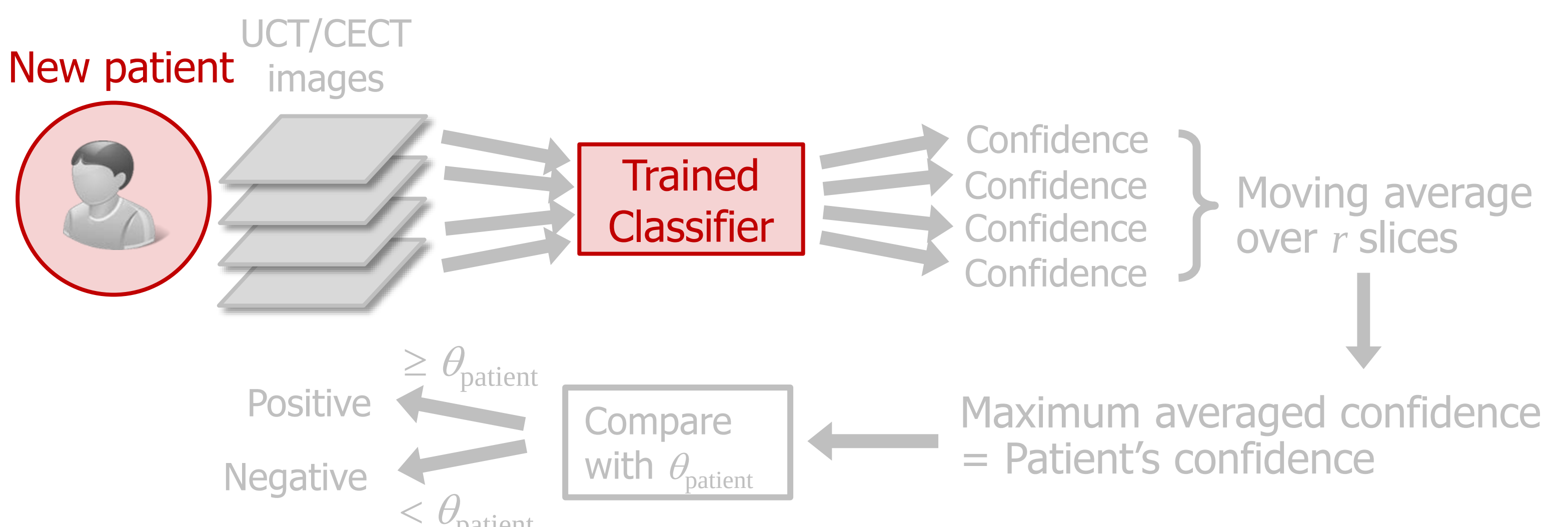
- ✓ Background
- ✓ Methods
- Experimental Results
- Conclusion

Results: Evaluation Metrics

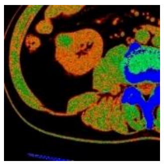
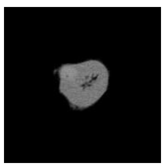
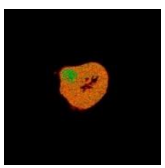
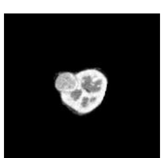
- We evaluated classifiers using Precision, Recall, F-measure, and AUROC of:
  - Image-wise detection to evaluate the classifiers directly



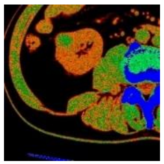
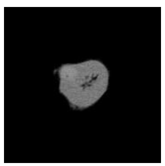
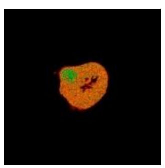
  - Patient-wise detection to evaluate the classifier from a practical viewpoint

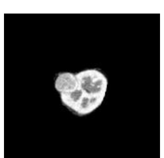


Results: Image-wise Detection

CT Images	Model	Precision	Recall	F-measure	AUROC
 UCT	VGG-16	0.374	0.174	0.237	0.679
	ResNet-50	0.280	0.398	0.328	0.713
	ViT	0.284	0.485	0.358	0.735
	SwinT	0.510	0.270	0.353	0.746
 UCT + 3ch	VGG-16	0.262	0.327	0.291	0.678
	ResNet-50	0.315	0.378	0.343	0.729
	ViT	0.344	0.367	0.356	0.749
	SwinT	0.417	0.357	0.384	0.747
 UCT + Mask	VGG-16	0.439	0.519	0.476	0.816
	ResNet-50	0.450	0.508	0.477	0.820
	ViT	0.449	0.567	0.501	0.825
	SwinT	0.586	0.476	0.525	0.854
 UCT + 3ch + Mask	VGG-16	0.484	0.567	0.522	0.846
	ResNet-50	0.453	0.540	0.493	0.818
	ViT	0.487	0.594	0.535	0.841
	SwinT	0.620	0.594	0.607	0.854
 CECT	VGG-16	0.500	0.464	0.482	0.818
	ResNet-50	0.528	0.538	0.529	0.823
	ViT	0.748	0.546	0.631	0.906
	SwinT	0.855	0.480	0.614	0.901
 CECT + Mask	VGG-16	0.711	0.500	0.587	0.845
	ResNet-50	0.638	0.589	0.612	0.853
	ViT	0.754	0.672	0.711	0.926
	SwinT	0.785	0.703	0.742	0.944

Results: Image-wise Detection

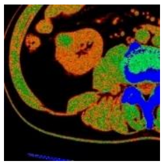
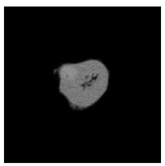
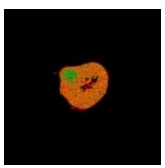





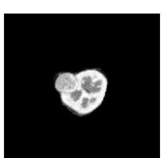


CT Images	Model	Precision	Recall	F-measure	AUROC
UCT	VGG-16	0.374	0.174	0.237	0.679
	ResNet-50	0.280	0.398	0.328	0.713
	ViT	0.284	0.485	0.358	0.735
	SwinT	0.510	0.270	0.353	0.746
UCT + 3ch	VGG-16	0.262	0.327	0.291	0.678
	ResNet-50	0.315	0.378	0.343	0.729
	ViT	0.344	0.367	0.356	0.749
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
UCT + Mask	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	SwinT	0.586	0.476	0.525	0.854
UCT + 3ch + Mask	VGG-16	0.484	0.567	0.522	0.846
	ResNet-50	0.453	0.540	0.493	0.818
	ViT	0.487	0.594	0.535	0.841
	SwinT	0.620	0.594	0.607	0.854
CECT	VGG-16	0.500	0.464	0.482	0.818
	ResNet-50	0.528	0.538	0.529	0.823
	ViT	0.748	0.546	0.631	0.906
	SwinT	0.855	0.480	0.614	0.901
CECT + Mask	VGG-16	0.711	0.500	0.587	0.845
	ResNet-50	0.638	0.589	0.612	0.853
	ViT	0.754	0.672	0.711	0.926
	SwinT	0.785	0.703	0.742	0.944

At the price of clinical costs for some patients, the accuracies for CECT images were higher than those for UCT images

Results: Image-wise Detection

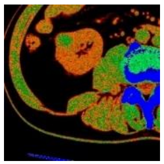
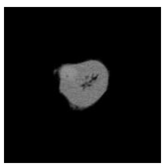
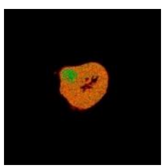
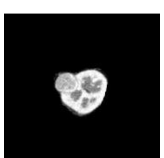







CT Images	Model	Precision	Recall	F-measure	AUROC
UCT	VGG-16	0.374	0.174	0.237	0.679
	ResNet-50	0.280	0.398	0.328	0.713
	ViT	0.284	0.485	0.358	0.735
	SwinT	0.510	0.270	0.353	0.746
UCT + 3ch	VGG-16	0.262	0.327	0.291	0.678
	ResNet-50	0.315	0.378	0.343	0.729
	ViT	0.344	0.367	0.356	0.749
	SwinT	0.417	0.357	0.384	0.747
UCT + Mask	VGG-16	0.439	0.519	0.476	0.816
	ResNet-50	0.450	0.508	0.477	0.820
	ViT	0.449	0.567	0.501	0.825
	SwinT	[illegible]	[illegible]	[illegible]	[illegible]
UCT + 3ch + Mask	VGG-16	[illegible]	[illegible]	[illegible]	[illegible]
	ResNet-50	[illegible]	[illegible]	[illegible]	[illegible]
	ViT	[illegible]	[illegible]	[illegible]	[illegible]
	SwinT	0.620	0.594	0.607	0.834
CECT	VGG-16	0.500	0.464	0.482	0.818
	ResNet-50	0.528	0.538	0.529	0.823
	ViT	0.748	0.546	0.631	0.906
	SwinT	0.855	0.480	0.614	0.901
CECT + Mask	VGG-16	0.711	0.500	0.587	0.845
	ResNet-50	0.638	0.589	0.612	0.853
	ViT	0.754	0.672	0.711	0.926
	SwinT	0.785	0.703	0.742	0.944

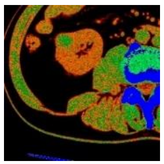
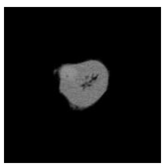
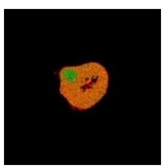
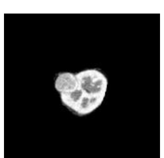
At the price of annotation costs,
the accuracies for masked images were
higher than those for unmasked images

Results: Image-wise Detection

CT Images	Model	Precision	Recall	F-measure	AUROC
 UCT	VGG-16	0.374	0.174	0.237	0.679
	ResNet-50	0.280	0.398	0.328	0.713
	ViT	0.284	0.485	0.358	0.735
	SwinT	0.510	0.270	0.353	0.746
 UCT + 3ch	VGG-16	0.262	0.327	0.291	0.678
	ResNet-50	0.315	0.378	0.343	0.729
	ViT	0.344	0.367	0.356	0.749
	SwinT	0.417	0.357	0.384	0.747
 UCT + Mask	VGG-16	0.439	0.519	0.476	0.816
	ResNet-50	0.450	0.508	0.477	0.820
	ViT	0.449	0.567	0.501	0.825
	SwinT	0.586	0.476	0.525	0.854
 UCT + 3ch + Mask	VGG-16	0.484	0.567	0.522	0.846
	ResNet-50	0.453	0.540	0.493	0.818
	ViT	0.487	0.594	0.535	0.841
	SwinT	0.620	0.594	0.607	0.854
 CECT	VGG-16	0.500	0.464	0.482	0.818
	ResNet-50	0.528	0.538	0.529	0.823
 CECT + Mask	ResNet-50	0.638	0.589	0.612	0.853
	ViT	0.754	0.672	0.711	0.926
	SwinT	0.785	0.703	0.742	0.944

Virtual colorization of UCT images was beneficial, considering its low execution cost

Results: Image-wise Detection

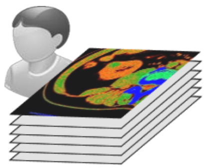
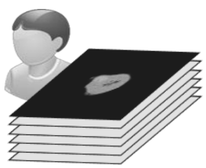
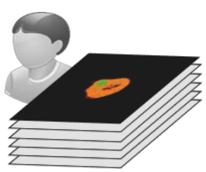
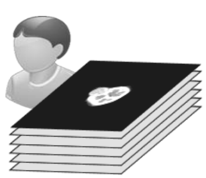
	CT Images	Model	Precision	Recall	F-measure	AUROC
	UCT	VGG-16	0.374	0.174	0.237	0.679
		ResNet-50	0.280	0.398	0.328	0.713
		ViT	0.284	0.485	0.358	0.735
		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	UCT + 3ch	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		ResNet-50	[illegible]	[illegible]	[illegible]	[illegible]
		ViT	0.344	0.367	0.356	0.749
		SwinT	0.417	0.357	0.384	0.747
	UCT + Mas[illegible]	VGG-16	0.439	0.519	0.476	0.816
		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	UCT + 3ch + Mas[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	CECT	VGG-16	[illegible]	[illegible]	[illegible]	[illegible]
		ResNet-50	0.528	0.538	0.529	0.823
		ViT	0.748	0.546	0.631	0.906
		SwinT	0.855	0.480	0.614	0.901
	CECT + Mask	VGG-16	0.711	0.500	0.587	0.845
		ResNet-50	0.638	0.589	0.612	0.853
		ViT	0.754	0.672	0.711	0.926
		SwinT	0.785	0.703	0.742	0.944

Transformer-based models generally performed better than CNNs

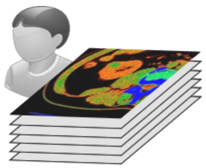
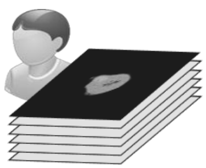
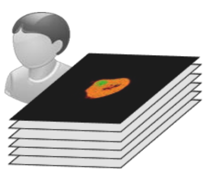
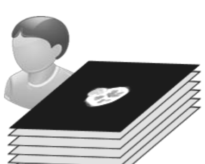
Difference turned to be larger for CECT images

Hypothesis:
Since ViT is biased towards shape [Turi+ 21], the shape of the contrast-enhanced part might have been better captured

Results: Patient-wise Detection

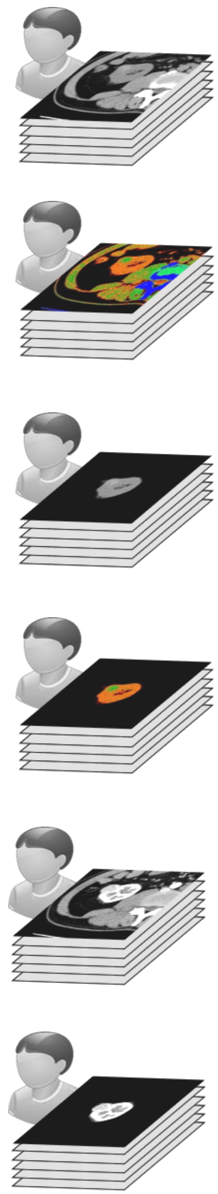
CT Images	Model	Precision	Recall	F-measure	AUROC
 UCT	VGG-16	0.540	0.692	0.607	0.596
	ResNet-50	0.491	0.718	0.583	0.608
	ViT	0.514	0.949	0.667	0.671
	SwinT	0.595	0.641	0.617	0.699
 UCT + 3ch	VGG-16	0.516	0.846	0.641	0.580
	ResNet-50	0.676	0.590	0.630	0.695
	ViT	0.634	0.667	0.650	0.682
	SwinT	0.608	0.795	0.689	0.701
 UCT + Mask	VGG-16	0.714	0.769	0.741	0.801
	ResNet-50	0.725	0.744	0.734	0.843
	ViT	0.844	0.692	0.761	0.799
	SwinT	0.833	0.769	0.800	0.848
 UCT + 3ch + Mask	VGG-16	0.682	0.769	0.723	0.798
	ResNet-50	0.623	0.846	0.717	0.837
	ViT	0.721	0.795	0.756	0.811
	SwinT	0.723	0.872	0.791	0.859
 CECT	VGG-16	0.688	0.564	0.620	0.733
	ResNet-50	0.735	0.641	0.685	0.762
	ViT	0.794	0.692	0.740	0.869
	SwinT	0.933	0.718	0.812	0.899
 CECT + Mask	VGG-16	0.674	0.846	0.750	0.844
	ResNet-50	0.725	0.744	0.734	0.822
	ViT	0.731	0.974	0.835	0.940
	SwinT	0.875	0.897	0.886	0.958

Results: Patient-wise Detection

CT Images	Model	Precision	Recall	F-measure	AUROC
 UCT	VGG-16	0.540	0.692	0.607	0.596
	ResNet-50	0.491	0.718	0.583	0.608
	ViT	0.514	0.949	0.667	0.671
	SwinT	0.595	0.641	0.617	0.699
 UCT + 3ch	VGG-16	0.516	0.846	0.641	0.580
	ResNet-50	0.676	0.590	0.630	0.695
	ViT	0.634	0.667	0.650	0.682
	SwinT	0.608	0.795	0.689	0.701
 UCT + Mask	VGG-16	0.714	0.769	0.741	0.801
	ResNet-50	0.725	0.744	0.734	0.843
	ViT	0.844	0.692	0.761	0.799
	SwinT	0.833	0.769	0.800	0.848
 UCT + 3ch + Mask	VGG-16	0.682	0.769	0.723	0.798
	ResNet-50	0.623	0.846	0.717	0.837
	ViT	0.721	0.795	0.756	0.811
	SwinT	0.723	0.872	0.791	0.859
 CECT	VGG-	[illegible]	[illegible]	[illegible]	[illegible]
	ResNe	[illegible]	[illegible]	[illegible]	[illegible]
	ViT	[illegible]	[illegible]	[illegible]	[illegible]
	SwinT	0.933	0.718	0.812	0.899
 CECT + Mask	VGG-16	0.674	0.846	0.750	0.844
	ResNet-50	0.725	0.744	0.734	0.822
	ViT	0.731	0.974	0.835	0.940
	SwinT	0.875	0.897	0.886	0.958

ViT did not perform better than ResNet-50 for UCT images

Results: Patient-wise Detection



CT Images	Model	Precision	Recall	F-measure	AUROC
UCT	VGG-16	0.540	0.692	0.607	0.596
	ResNet-50	0.491	0.718	0.583	0.608
	ViT	0.514	0.949	0.667	0.671
	SwinT	0.595	0.641	0.617	0.699
UCT + 3ch	VGG-16	0.516	0.846	0.641	0.580
	ResNet-50	[illegible]	[illegible]	[illegible]	[illegible]
	ViT	0.634	0.667	0.650	0.682
	SwinT	0.608	0.795	0.689	0.701
UCT + Mask	VGG-16	0.714	0.769	0.741	0.801
	ResNet-50	0.725	0.744	0.734	0.843
	ViT	0.844	0.692	0.761	0.799
	SwinT	0.833	0.769	0.800	0.848
UCT + 3ch + Mask	VGG-16	0.682	0.769	0.723	0.798
	ResNet-50	0.623	0.846	0.717	0.837
	ViT	0.721	0.795	0.756	0.811
	SwinT	0.723	0.872	0.791	0.859
CECT	VGG-16	0.688	0.564	0.620	0.733
	ResNet-50	0.735	0.641	0.685	0.762
	ViT	0.794	0.692	0.740	0.869
	SwinT	0.933	0.718	0.812	0.899
CECT + Mask	VGG-16	0.674	0.846	0.750	0.844
	ResNet-50	0.725	0.744	0.734	0.822
	ViT	0.731	0.974	0.835	0.940
	SwinT	0.875	0.897	0.886	0.958

Swin Transformer generally worked the best

Outline

- ✓ Background
- ✓ Methods
- ✓ Experimental Results
- Conclusion

Conclusion

- We studied on detection of kidney cancer from CT images for 400+ (virtual) patients
- We examined CNN-based and Transformer-based classifiers (VGG-16, ResNet-50, ViT and Swin Transformer)
- We evaluated the accuracy across various types of CT images:
  - UCT images vs. CECT images
  - Masked vs. Unmasked
  - Grayscale vs. Virtually Colored
- Observations:
  - Predictive performance varied drastically depending on image types and preprocessings
  - Swin Transformer generally worked the best
  - Transformer-based models were effective especially for CECT images

Future Work

- Comparison with:
  - CNN-based models (e.g. EfficientNetV2 [Tan+ 21])
  - MLP-based models (e.g. MLP-Mixer [Tolstikhin+ 21])
  - CNN-Transformer hybrids (e.g. CvT [Wu+ 21])
- Introducing visual explanations methods (e.g. Transformer Explainability [Chefer+ 21])
- Coping with unavailability of CECT images:
  - Synthetic CECT images created by image-to-image models [Hu+ 21][Sassa+ 22]

Thank You for Your Attention!