Attention Based Cell Detection Enhancer

2024.02.28 Wed

Jeongin Park

Intern, Model-Centric AIR 1, Oncology, Lunit

2023.12.26 ~ 2024.02.29

Lunit

Goal of this Internship

Attention Based Cell Detection Enhancer

Jeongin Park

Intern, Model-Centric AIR 1, Oncology, Lunit

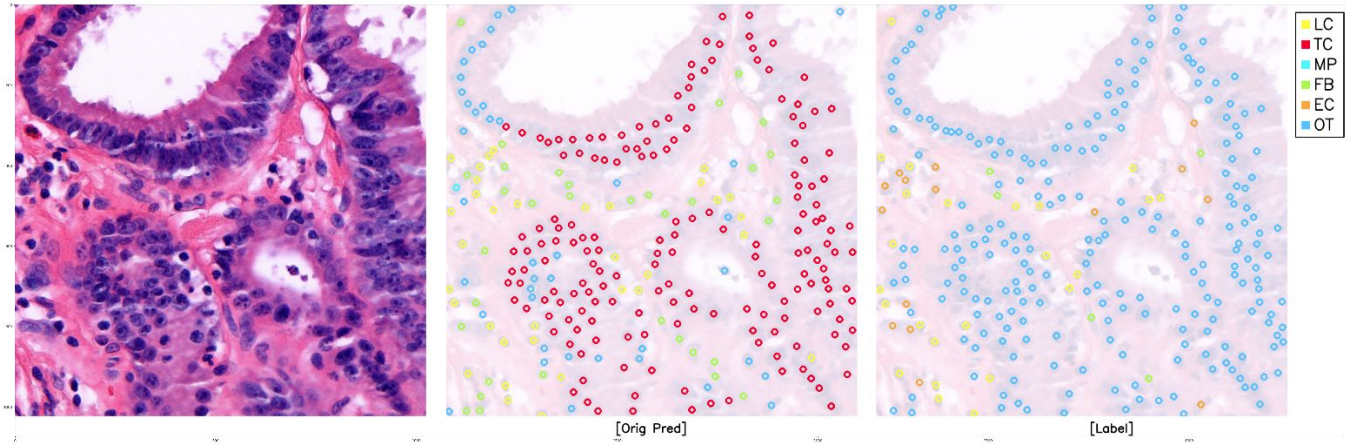
2023.12.26 ~ 2024.02.29

Lunit

Internship Goal

DIB-SCOPE Result

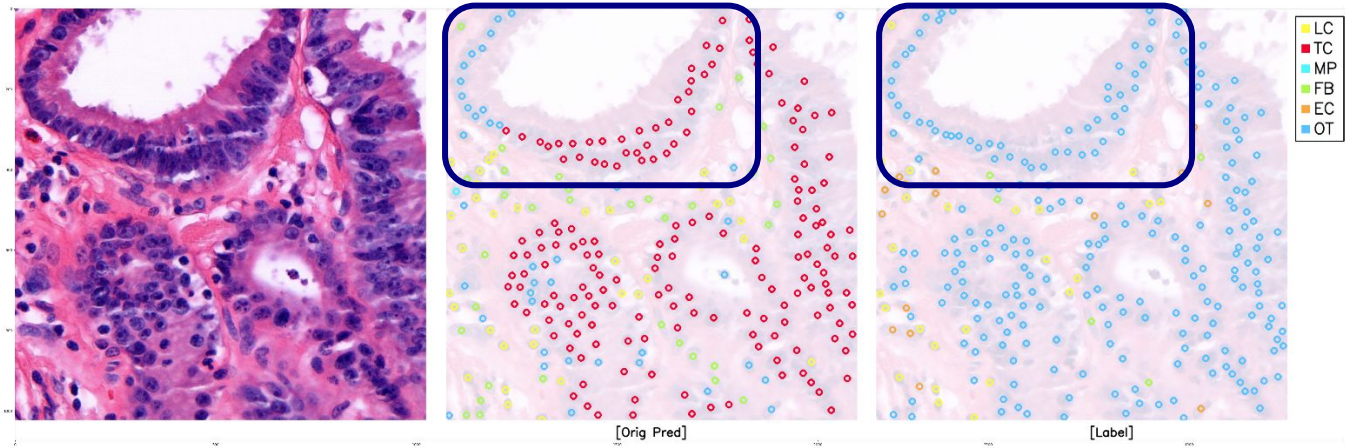
Original model predicts each point individually



Internship Goal

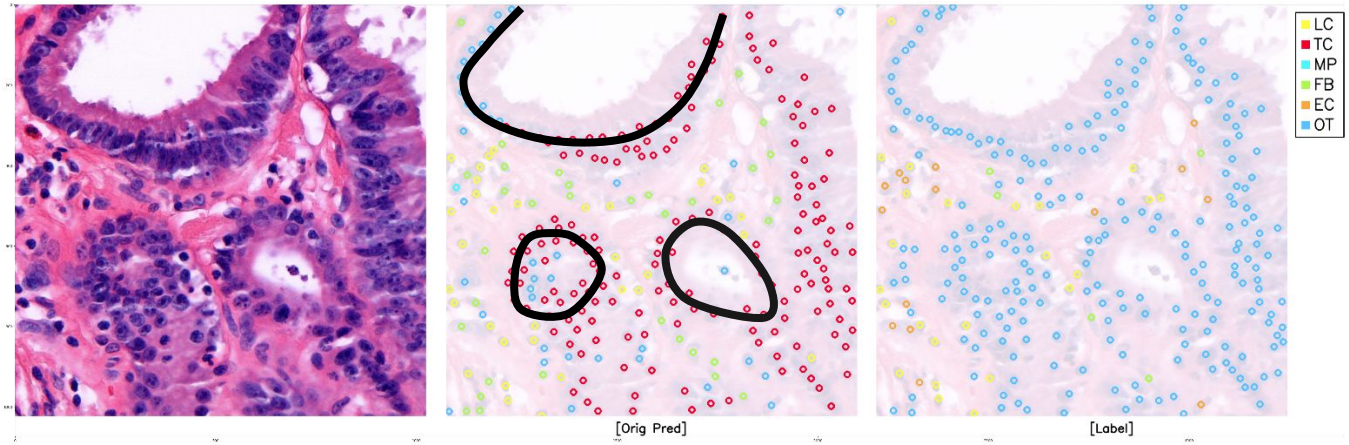
DIB-SCOPE Result

But it seems that they should also consider **relationship with other cells**



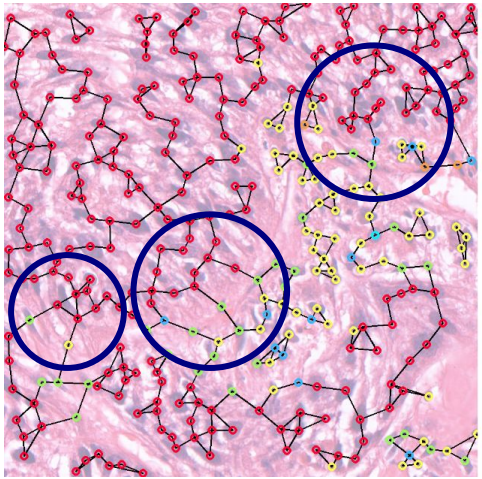
Idea Development

Can’t we draw some graphs?

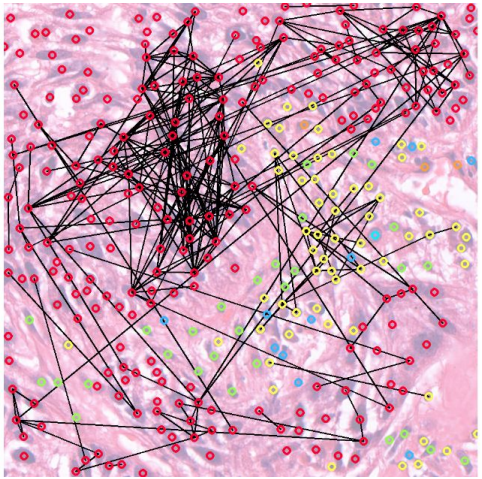


Idea Development

Draw Graphs with kNN and Feature map



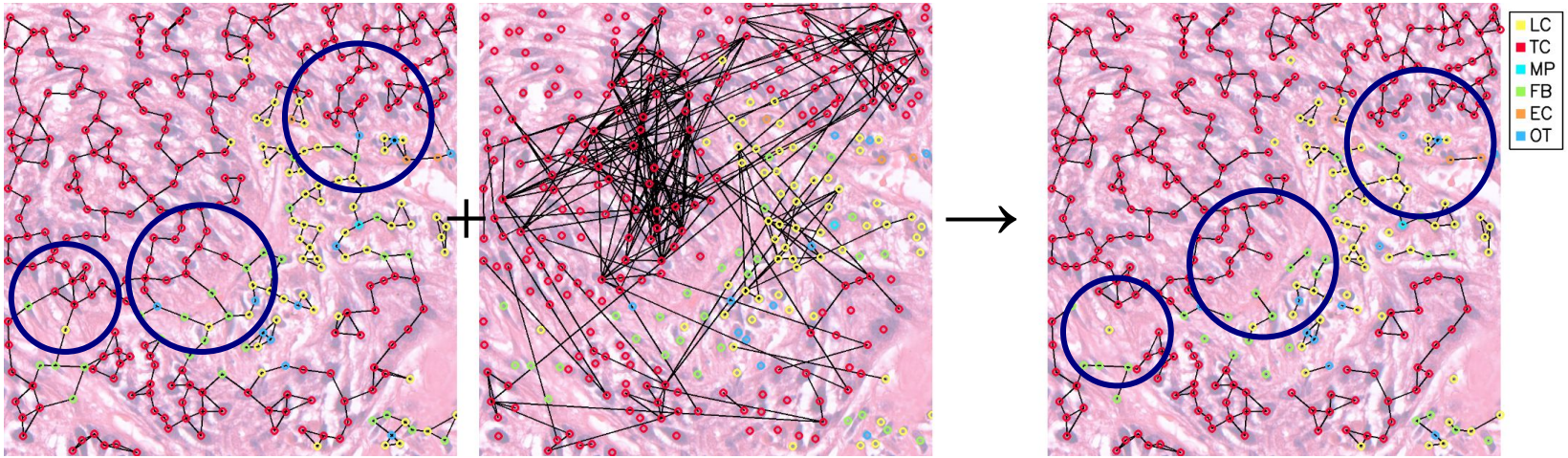
kNN
(k=3)



feature similarity
(threshold=0.97)

Idea Development

Draw Graphs with kNN and Feature map



kNN
(k=3)

feature similarity
(threshold=0.97)

kNN + feature similarity
(threshold= 0.8)

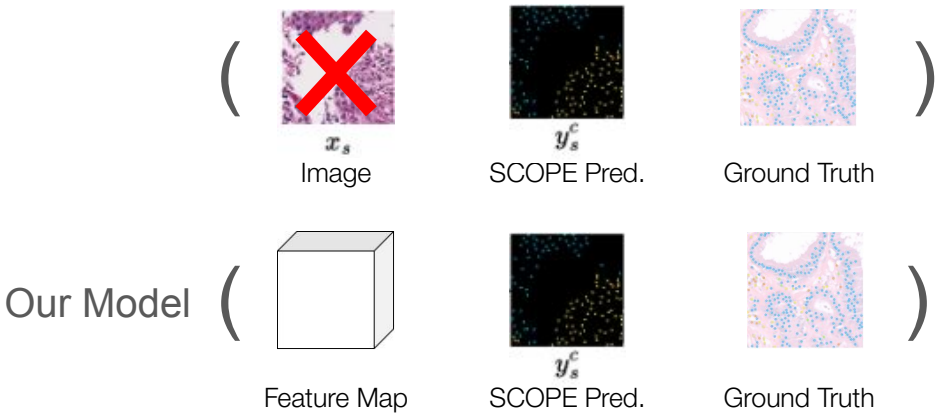
Considering both **position** (with kNN) and **feature** makes more reasonable result

Define the Problem

Attention Based Cell Detection Enhancer

Goal: to enhance cell detection result **by considering relationship between other cells**

How: use **attention** to utilize both **1. positional information** and **2. feature similarity**



Why not end-to-end? because of limited time, for Proof of Concept

How do we Implement our Idea

Attention Based Cell Detection Enhancer

Jeongin Park

Intern, Model-Centric AIR 1, Oncology, Lunit

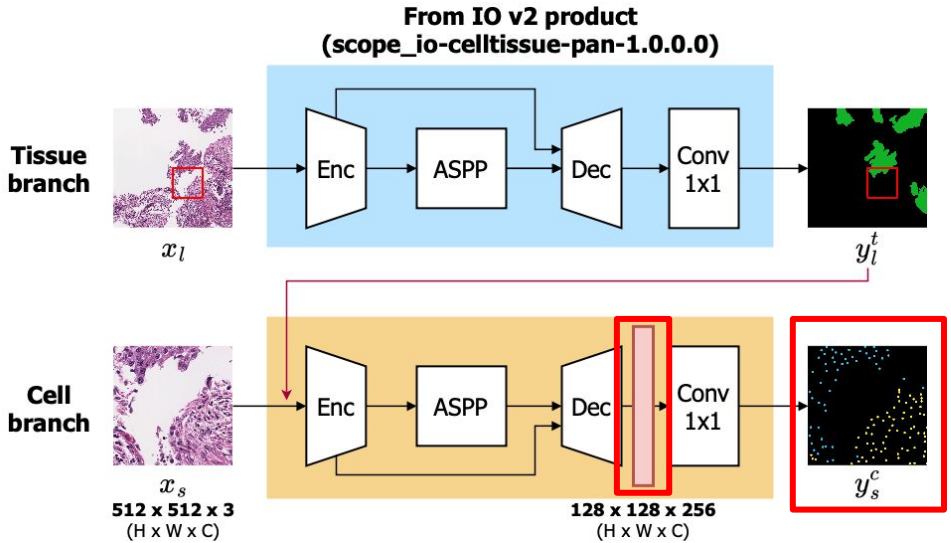
2023.12.26 ~ 2024.02.29

Lunit

Where do we get input

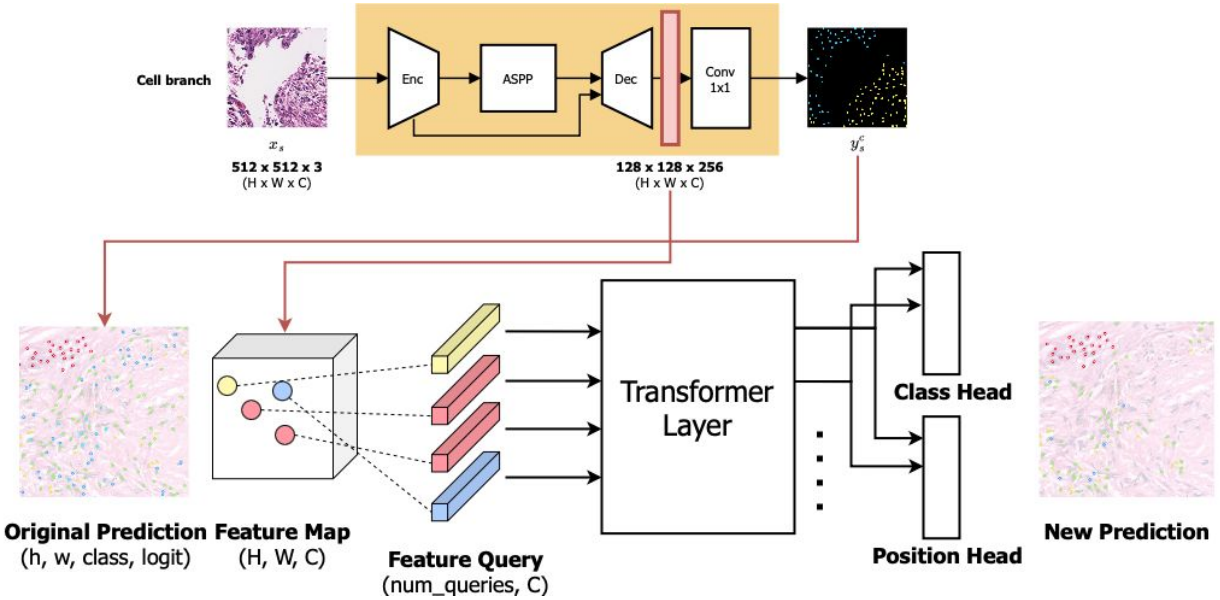
DIB-SCOPE Model

Inputs: **feature map** from DIB-SCOPE model and its **prediction result**



New Model Architecture

Use Transformer



Why do we use transformer? to utilize both **1. positional information** and **2. feature similarity**

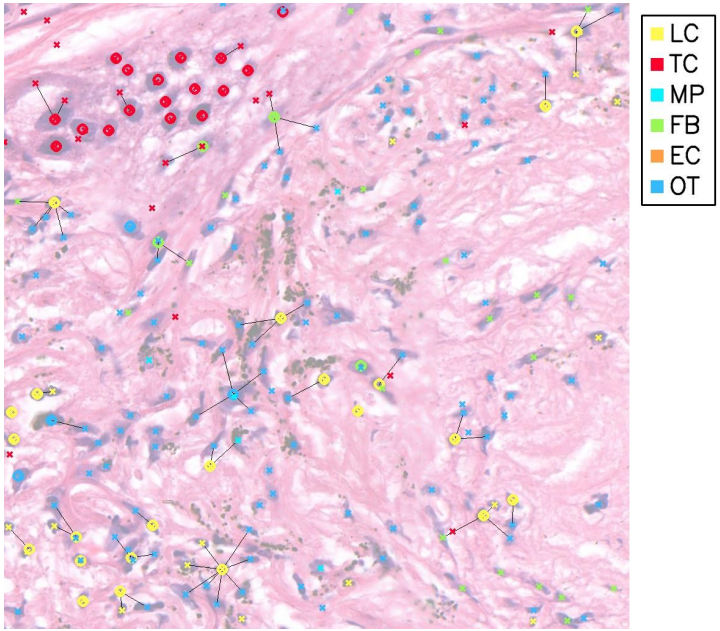
Bipartite Matching Loss

1. Match prediction and label points using Hungarian matcher
2. Calculate L1 loss (position) and cross-entropy loss (class)

$$\mathcal{L}(y,\,\hat{y}) = \sum_{i=1}^{N}\Big[-\log \hat{p}_{\hat{\sigma(i)}}(c_i) + \lambda\Big|\Big|y_i - \hat{y}_{\hat{\sigma(i)}}\Big|\Big|_1\Big]$$

$\hat{\sigma(i)}$: optimal permutaion from matching step

$\hat{p}_{\hat{\sigma(i)}}(c_i)$: probability of class c_i



Example of Matching Plot
x: Prediction, o: Label

Development of Model Architecture

Attention Based Cell Detection Enhancer

Jeongin Park

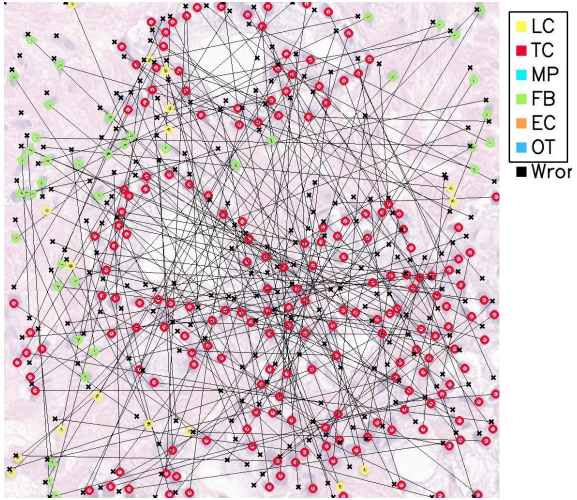
Intern, Model-Centric AIR 1, Oncology, Lunit

2023.12.26 ~ 2024.02.29

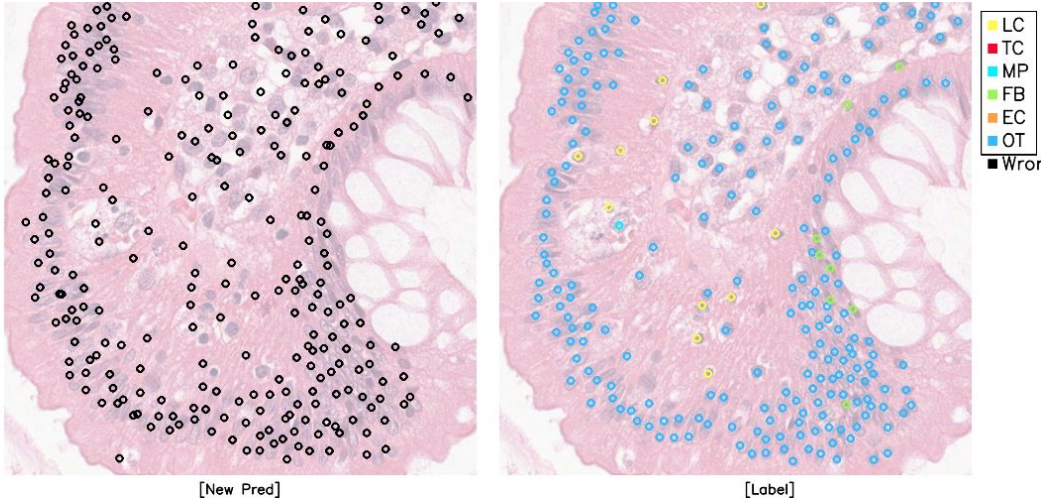
Lunit

Trials and Errors

Some minor issues



Due to no Normalization, wrong matching



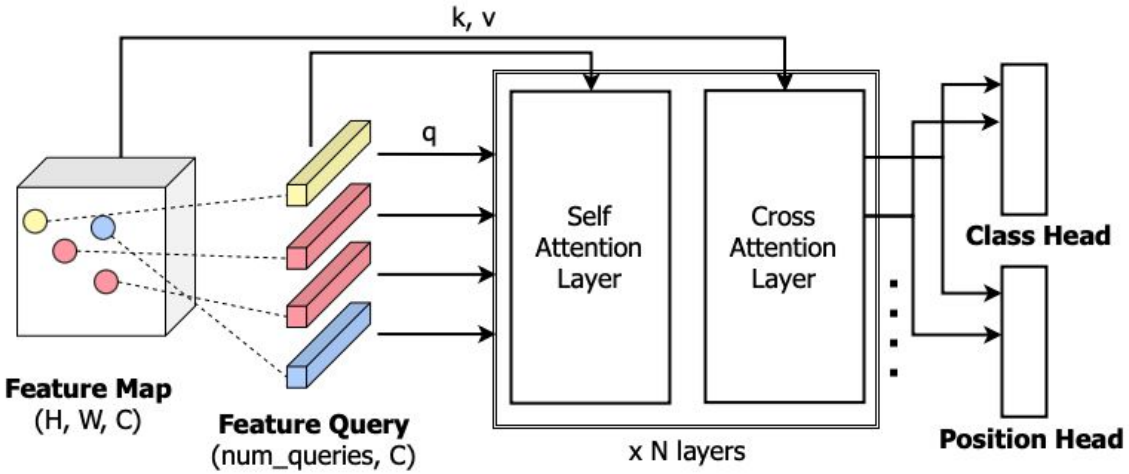
Wrong class label

Minimize Model Architecture

First model

Memory issue with GPU

Complexity: original cross attention layer: $O(N_q HW)$ $(H: 128, W: 128)$

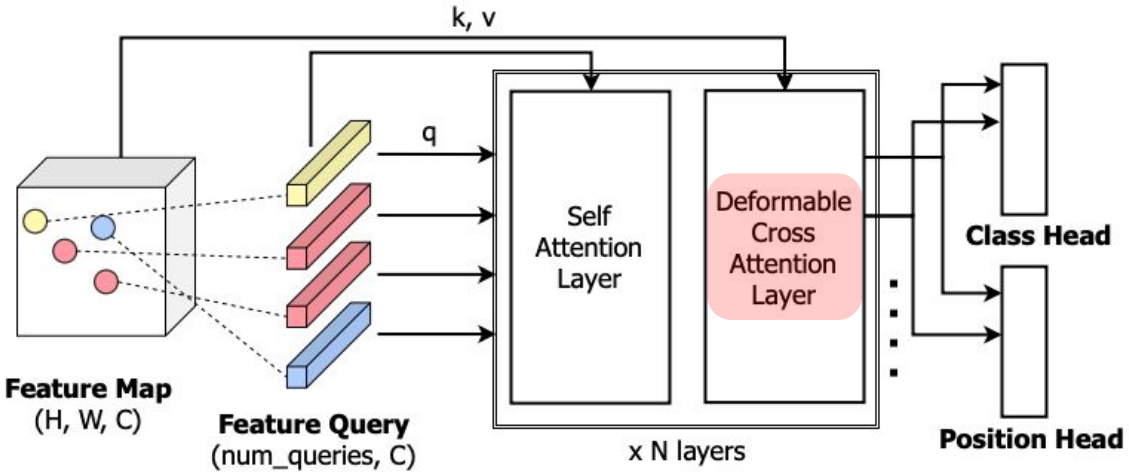


Minimize Model Architecture

Use Deformable Attention Layer

Memory issue with GPU → used **deformable cross attention layer** to select reference points

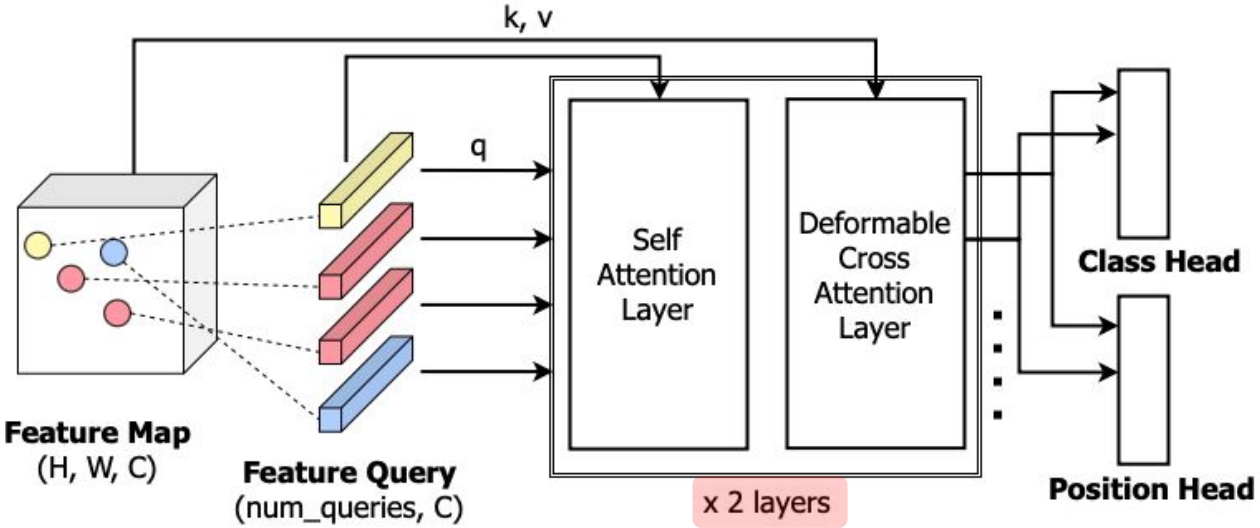
Complexity: original cross attention layer: $O(N_q HW)$ → deformable cross attention: $O(N_q K)$ $(K \ll HW)$



Minimize Model Architecture

Decrease number of Layers

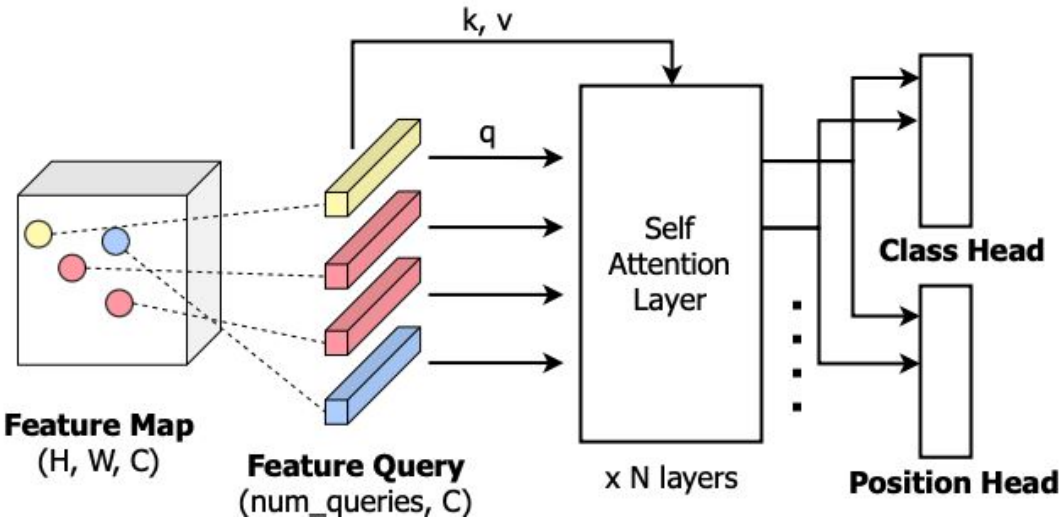
Model not being trained → decrease transformer layer from 6 to 2



Minimize Model Architecture

Only use Self Attention Layer

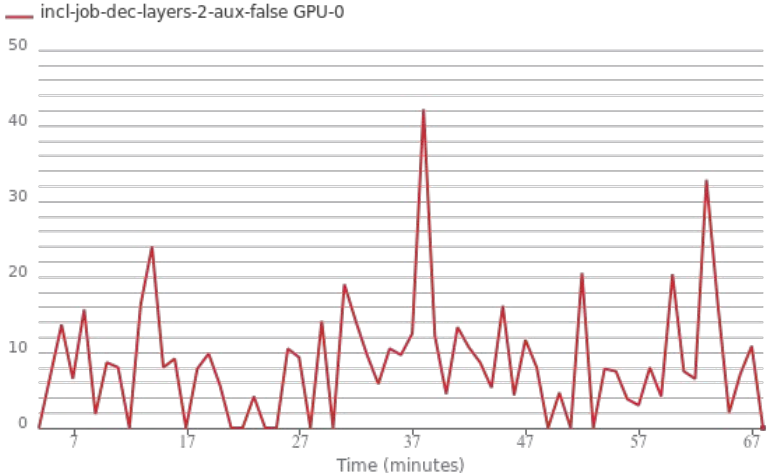
Transformer has too much capacity → Remove cross attention layer and leave only self attention layer



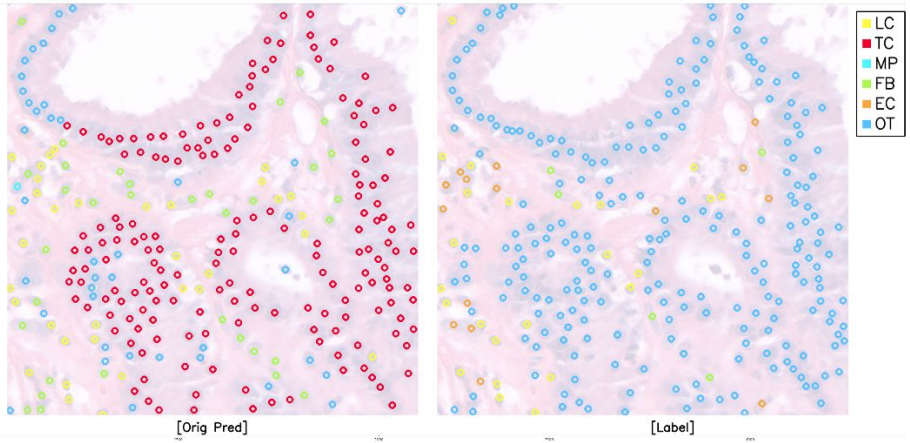
Predict only Class Label

Use original result of points

1. Issues: low GPU utilization, long training time, NaN issue with transformer mask → because of **matching**
2. Prediction of position was already good



GPU utilization per minute



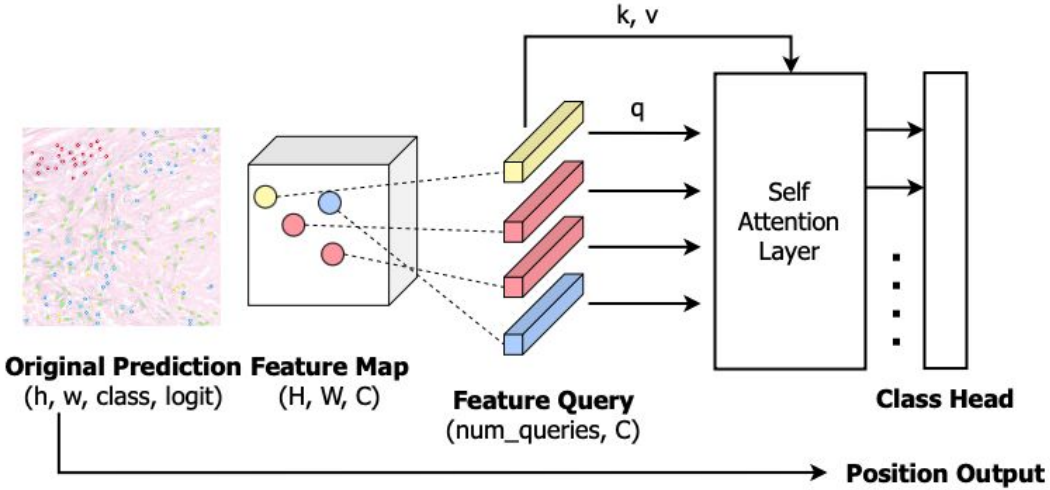
Prediction of original model and Ground Truth

Predict only Class Label

Use original result of points

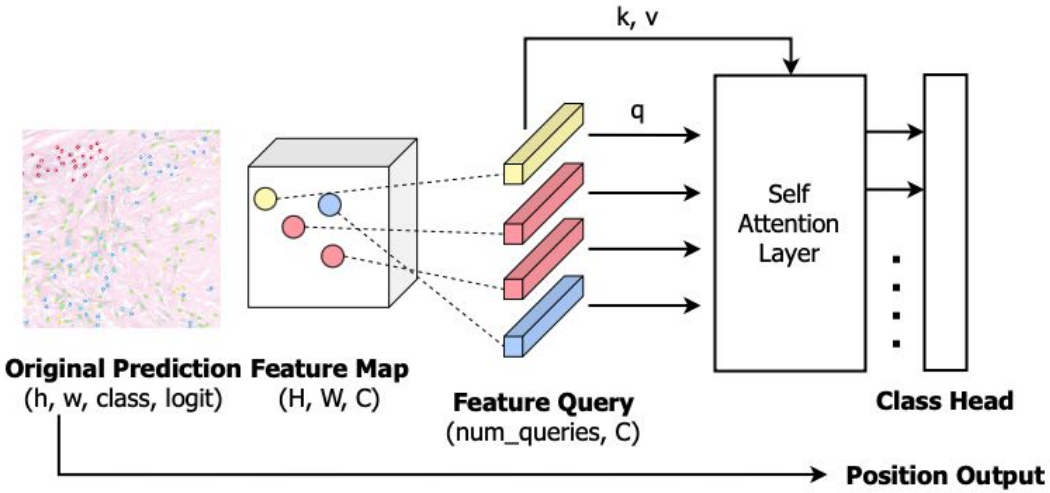
1. Issues: low GPU utilization, long training time, NaN issue with transformer mask
2. Prediction of position was already good

→ Match prediction and GTs in advance, and **predict only class**



Summary

Proceed with simple model and method, use more complex, bigger model later



Experiment Results

Attention Based Cell Detection Enhancer

Jeongin Park

Intern, Model-Centric AIR 1, Oncology, Lunit

2023.12.26 ~ 2024.02.29

Lunit

Experiment Results

Compared the results with different number of layers, optimizers, scheduler, etc.

But model did not improve from original score

Experiment Log here

	mF1	LC	TC	MP	FB	EC	OT	best epoch
Original	**0.5225**	**0.6944**	**0.7442**	**0.4099**	**0.3969**	**0.4386**	**0.4507**	X
New Model	**0.5169**	**0.6850**	**0.7430**	**0.4102**	**0.3579**	**0.4259**	**0.4791**	**12**

Validation set Result

	mF1	LC	TC	MP	FB	EC	OT	best epoch
Original Model	**0.5298**	**0.6992**	**0.8044**	**0.3123**	**0.5564**	**0.3886**	**0.4176**	X
New Model	**0.5178**	**0.6922**	**0.8045**	**0.3108**	**0.4855**	**0.3692**	**0.4448**	**12**

Test set Result

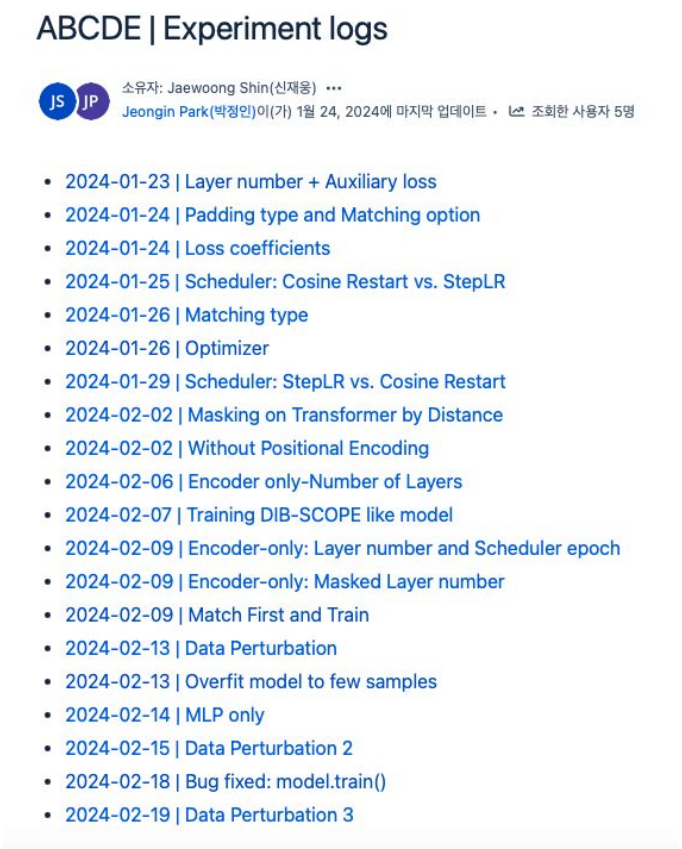
Experiment Results

Experiment Logs

Also more experiments with padding(zero or random padding),

which method of matching(Hungarian or greedy) to use, etc.

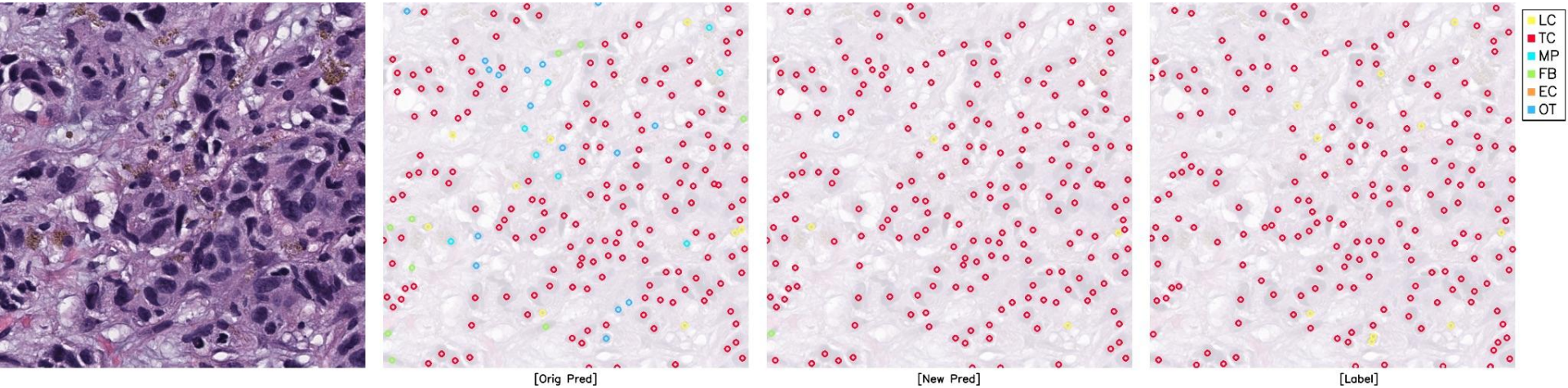
Detailed results can be found here



Visualization

Comparison with Original Result

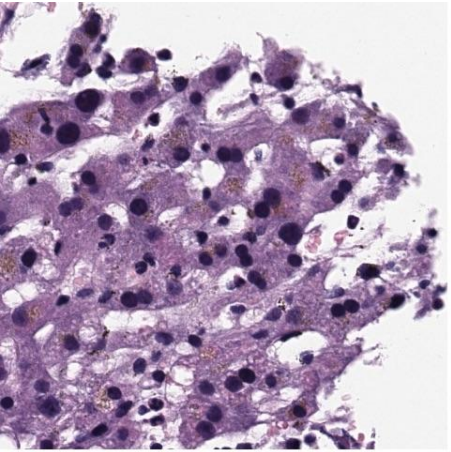
More accurate prediction of class

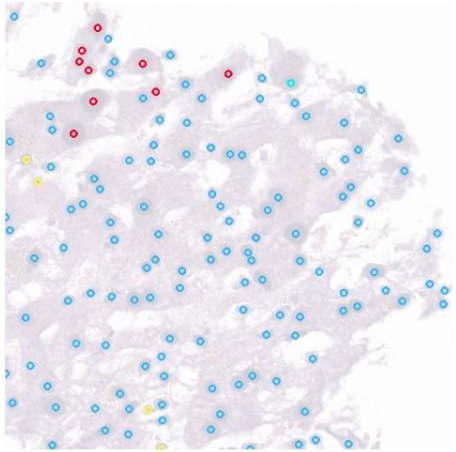


Visualization

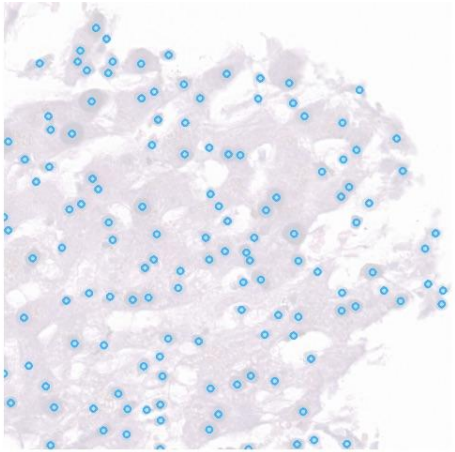
Comparison with Original Result

More accurate prediction of class

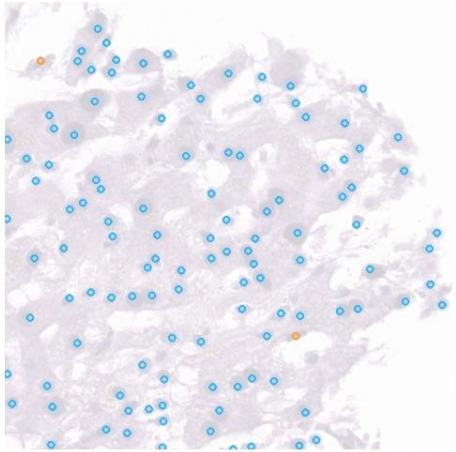




[Orig Pred]



[New Pred]

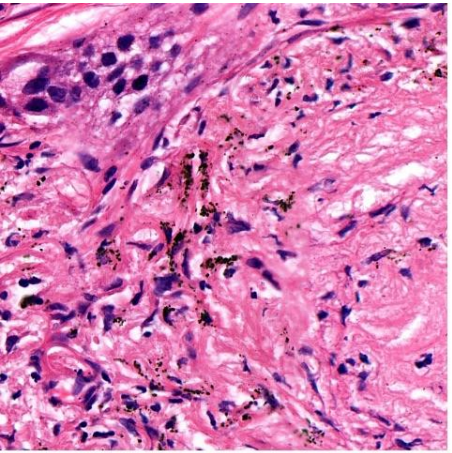


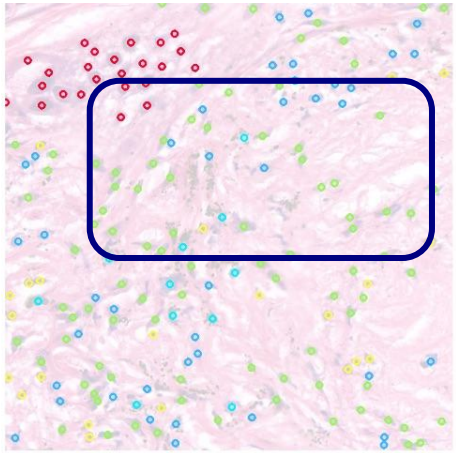
[Label]

Visualization

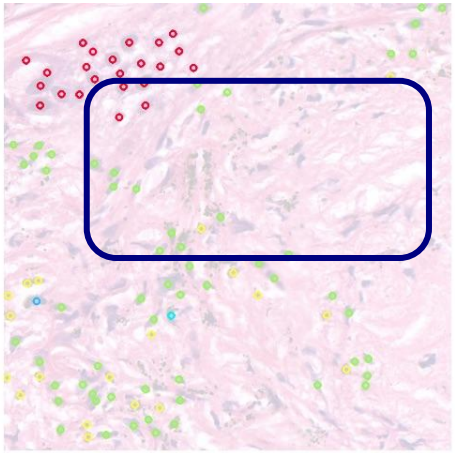
Comparison with Original Result

Erase False Positives

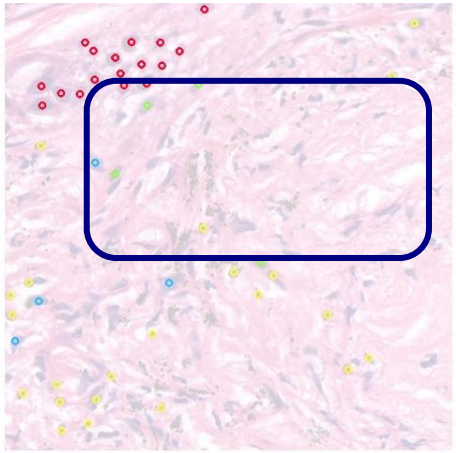




[Orig Pred]



[New Pred]

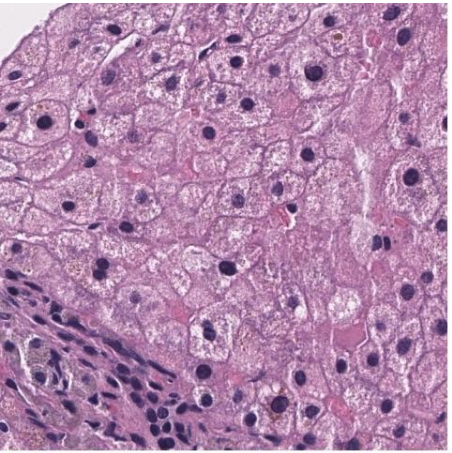


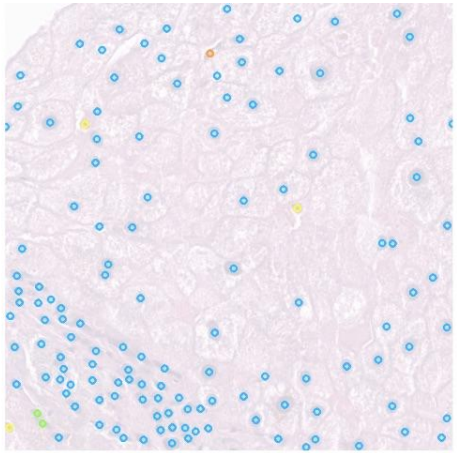
[Label]

Visualization

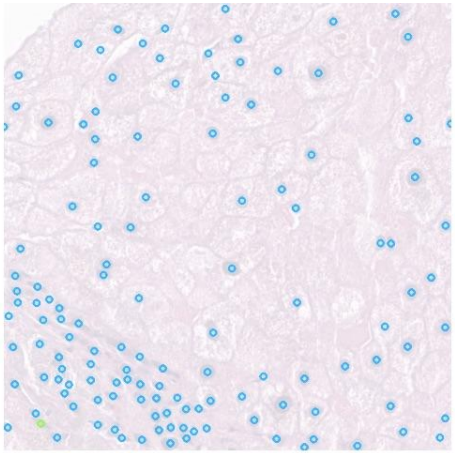
Comparison with Original Result

But some did not change a lot from original result

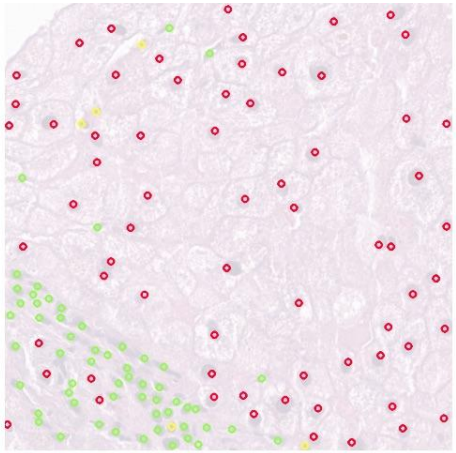




[Orig Pred]



[New Pred]



[Label]

Attention Map

Attention Based Cell Detection Enhancer

Jeongin Park

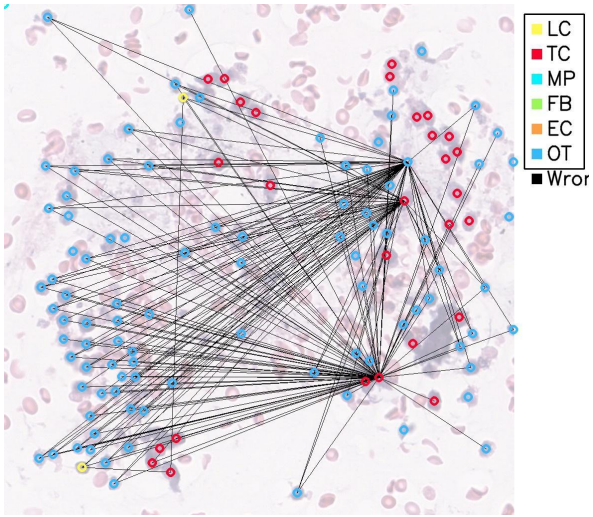
Intern, Model-Centric AIR 1, Oncology, Lunit

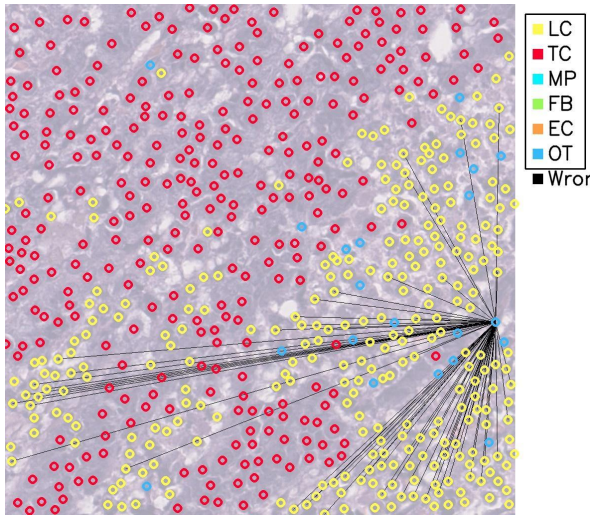
2023.12.26 ~ 2024.02.29

Lunit

Attention Map

Few points have strong attention





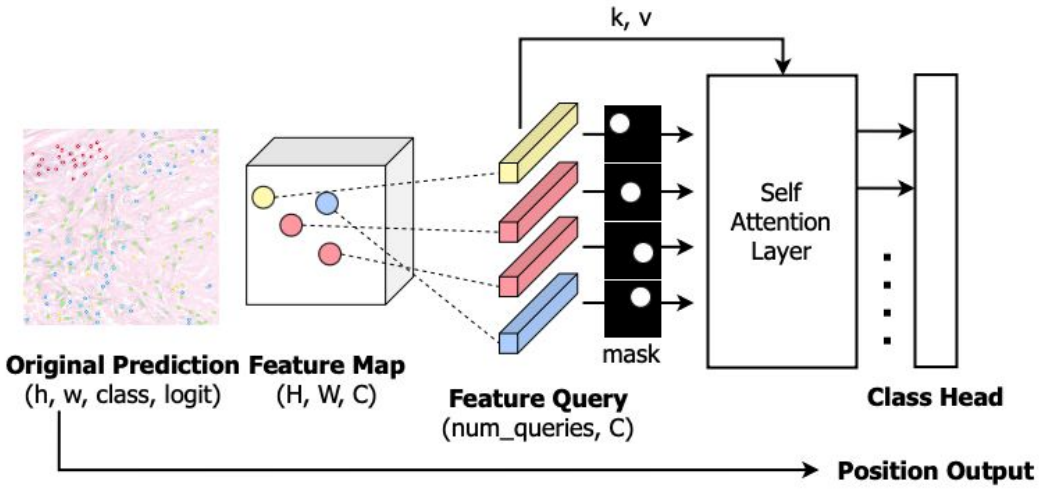
Line between points with high attention (top 50)

Attention Map

How about Adding Attention Mask according to Distance?

To focus more on local, topologically close points, we added mask to look only at close points

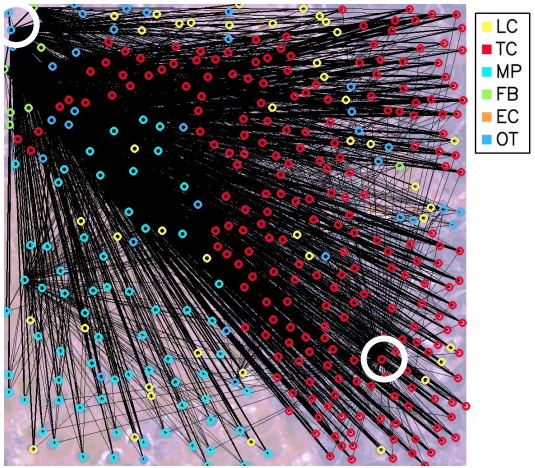
Experiment log with different masking distances can be found here

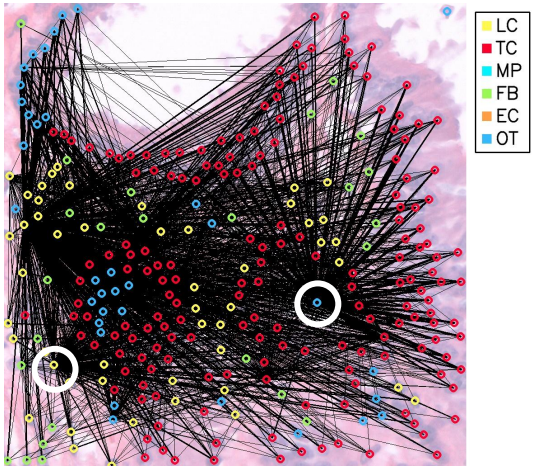


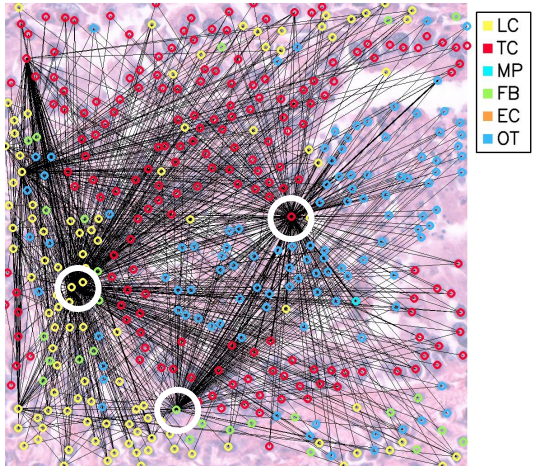
Attention Map

Before adding Attention Mask

Let’s add mask to look only at close (distance < 128) points





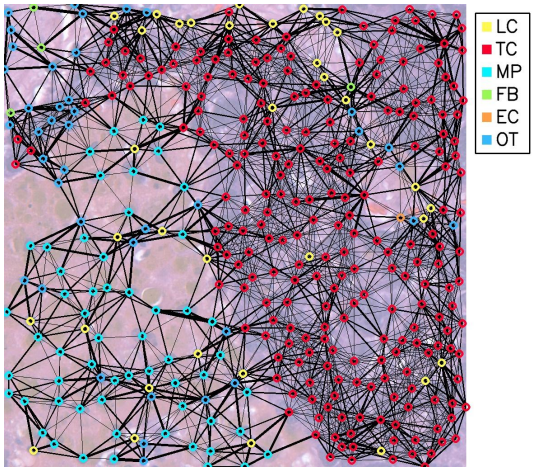


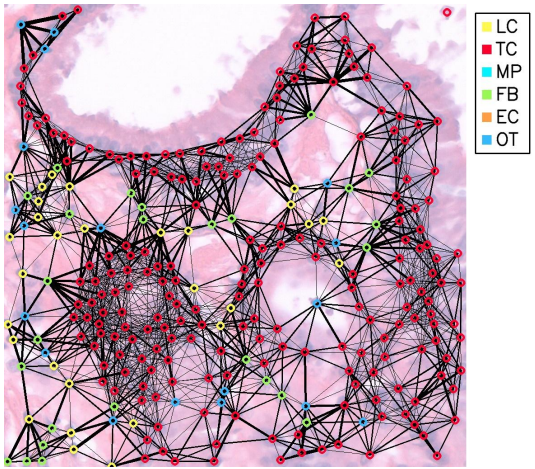
Line between points with high attention: threshold=0.02
(Thick lines meaning high attention)

Attention Map

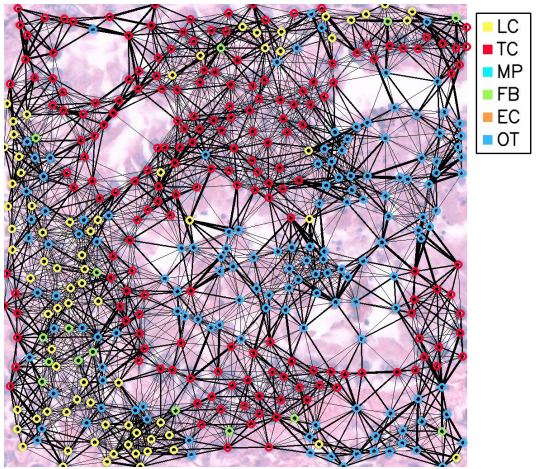
After adding Attention Mask

Adding mask to look only at close points





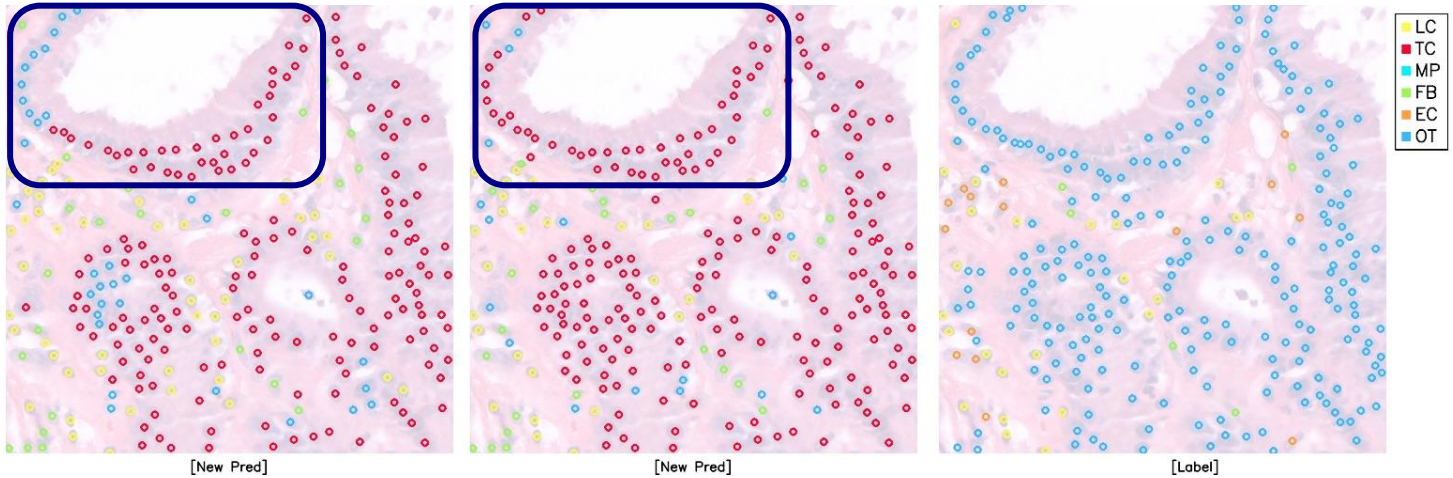
Line between all points
(Thick lines meaning high attention)



Attention Map

Effect of Attention Map

With Attention mask, cells that are close interact with each other



Without Masking, With Masking(128), and Ground Truth

Further Investigation

Attention Based Cell Detection Enhancer

Jeongin Park

Intern, Model-Centric AIR 1, Oncology, Lunit

2023.12.26 ~ 2024.02.29

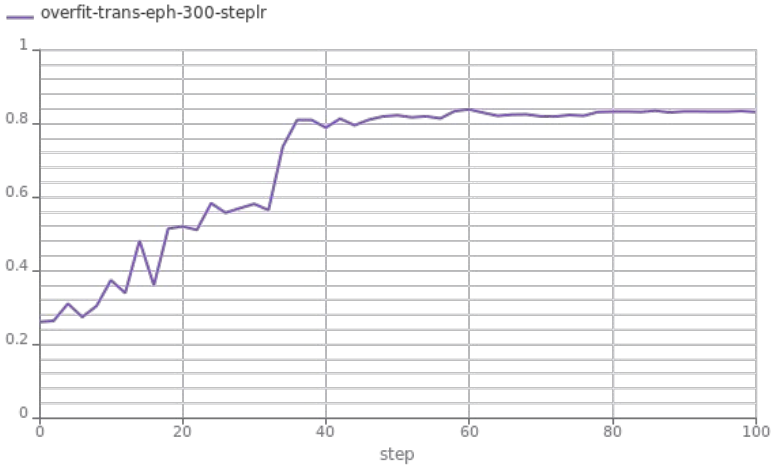
Lunit

Overfit to Few Samples

To check if model has possibility of being trained

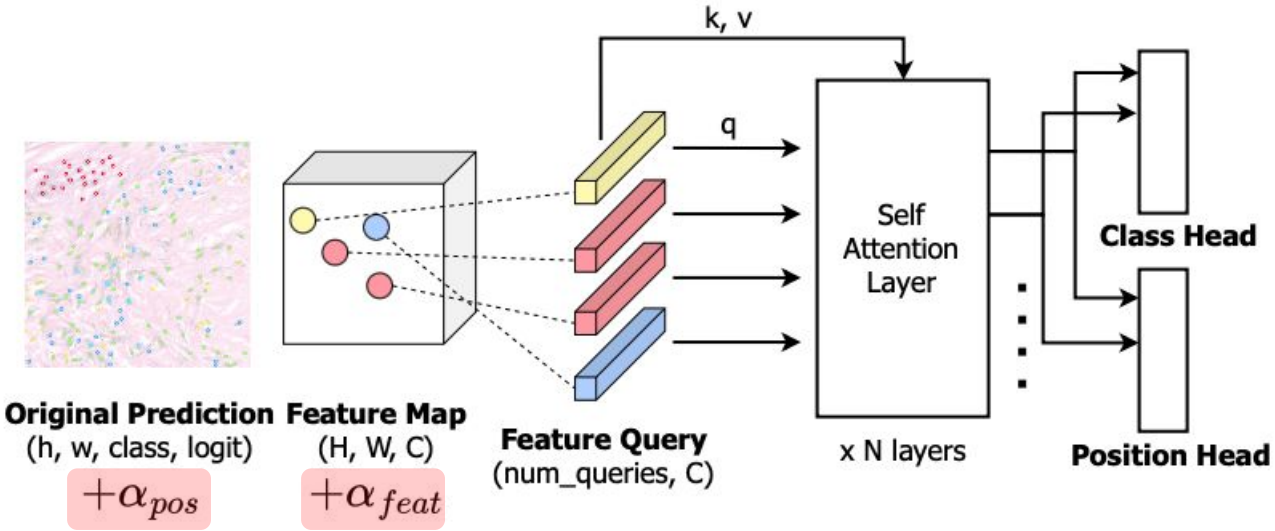
It takes at least 40 epoch to overfit to 8 samples

This suggests that it might take long time to train



Data Perturbation

Add noise to Feature map and Positional Information



Data Perturbation

Result

Through validation loss graph, we can conclude that if none or little noise is added, model overfit to training data

Thus, adding noise to feature map is effective to improve performance

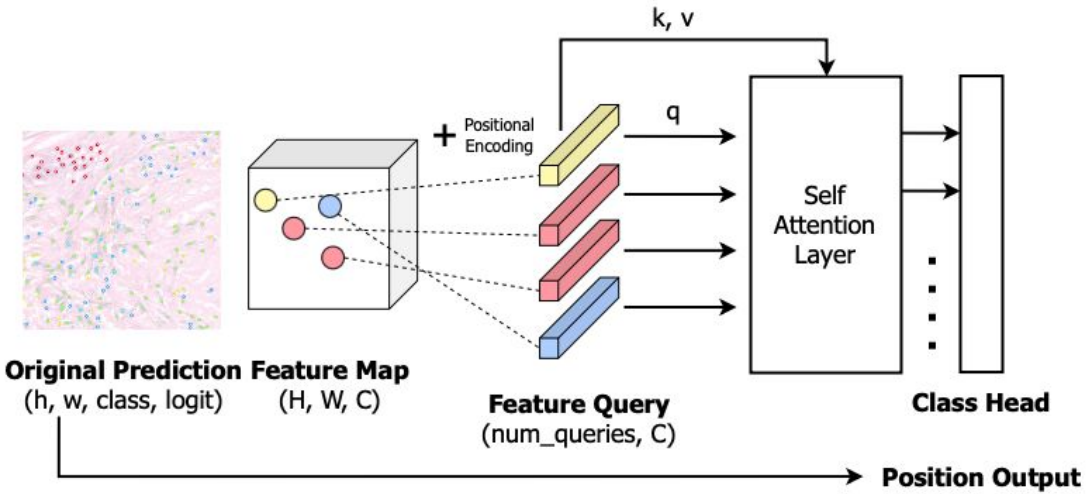
	mF1	LC	TC	MP	FB	EC	OT
Original Model	**0.5225**	**0.6944**	**0.7442**	**0.4099**	**0.3969**	**0.4386**	**0.4507**
0	0.5118	0.6676	0.7455	0.4262	0.3839	0.4336	0.4141
1e-2	0.5105	0.6632	0.7416	0.4221	0.3932	0.4034	0.4398
1e-3	0.5151	0.6679	0.7441	0.4208	0.3886	0.4273	0.4421
1	0.5175	0.6659	0.7440	0.4254	0.3894	0.4289	0.4513

Validation: Best mF1 Score

Relative Positional Encoding

Used Sinusoidal Positional Encoding before

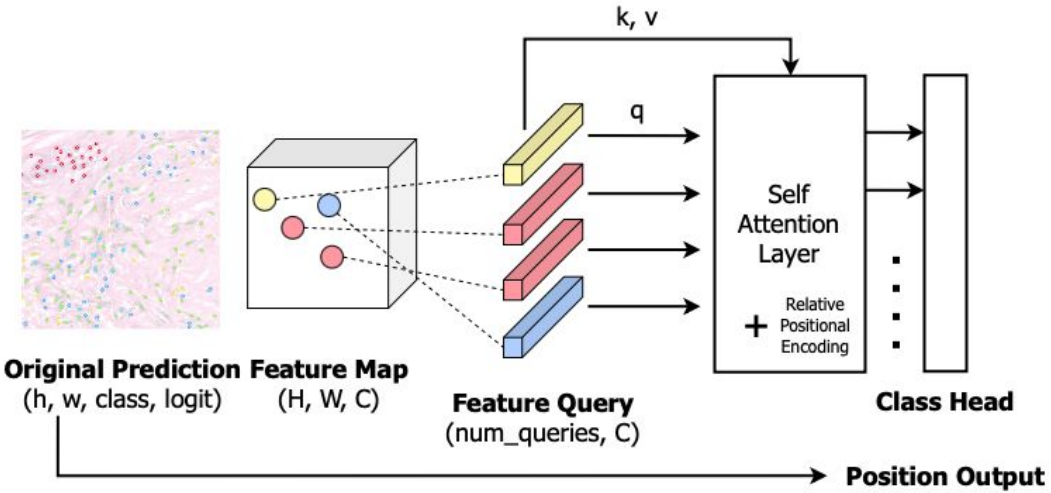
Former attention layer uses **sinusoidal positional encoding** according to original prediction of position



Relative Positional Encoding

L2 distance between queries

New method adds **relative positional encoding**, which is L2 distance between each query



$$A = \exp\left(QK^T + \Omega(m-n)\right)$$

$\Omega(m-n)$: Relative Positional Encoding
L2 distance between each query

Relative Positional Encoding

Result

Using learnable coefficient for relative positional encoding works best

	mF1	LC	TC	MP	FB	EC	OT
Baseline	**0.5263**	**0.6908**	**0.7474**	**0.4147**	**0.4068**	**0.4413**	**0.4567**
Sinusoidal	0.5175	0.6659	0.7440	0.4254	0.3894	0.4289	0.4513
Relative	0.5188	0.6738	0.7463	0.4256	0.3868	0.4364	0.4437
Learnable Relative	**0.5236**	**0.6752**	**0.7444**	**0.4287**	**0.3948**	**0.4326**	**0.4659**
Both	0.5184	0.6701	0.7458	0.4252	0.3780	0.4281	0.4629

Validation: Best mF1 Score

Relative Positional Encoding

Result

Using learnable relative positional encoding works best

	mF1	LC	TC	MP	FB	EC	OT
Baseline	**0.5298**	**0.6992**	**0.8044**	**0.3123**	**0.5564**	**0.3886**	**0.4176**
Sinusoidal	0.5028	0.6855	0.7987	0.2707	0.5387	0.3586	0.3645
Relative	0.5044	0.6818	0.8016	0.2790	0.5276	0.3632	0.3732
Learnable Relative	**0.5046**	**0.6843**	**0.7945**	**0.2661**	**0.5421**	**0.3724**	**0.3681**
Both	0.4998	0.6798	0.7975	0.2730	0.5009	0.3578	0.3896

Validation: Best mF1 Score

Summary

- Overfit to few samples
- Data perturbation
- Relative positional encoding

Summary

- Overfit to few samples → Model can learn!
- Data perturbation → necessary to avoid overfitting
- Relative positional encoding → works better than sinusoidal positional encoding

Conclusion

Attention Based Cell Detection Enhancer

Jeongin Park

Intern, Model-centric AIR 1, Oncology, Lunit

2023.12.26 ~ 2024.02.29

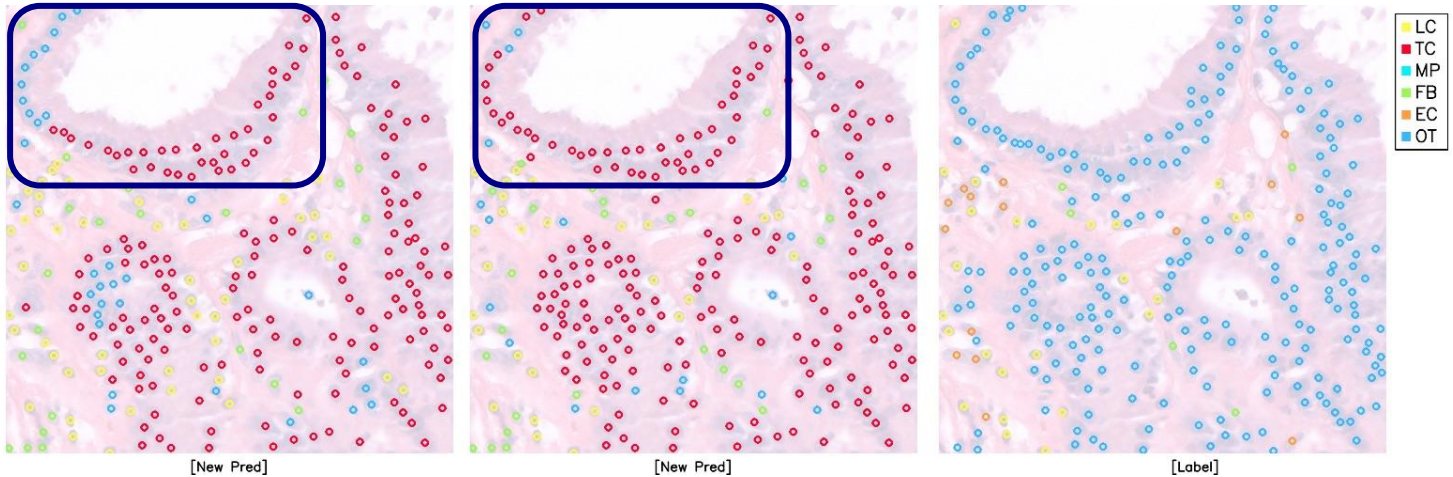
Lunit

Conclusion

Effect of Attention Map

Transformer can consider relationship between cells as seen in attention map

Even though performance score was not good, it has possibility to be well utilized

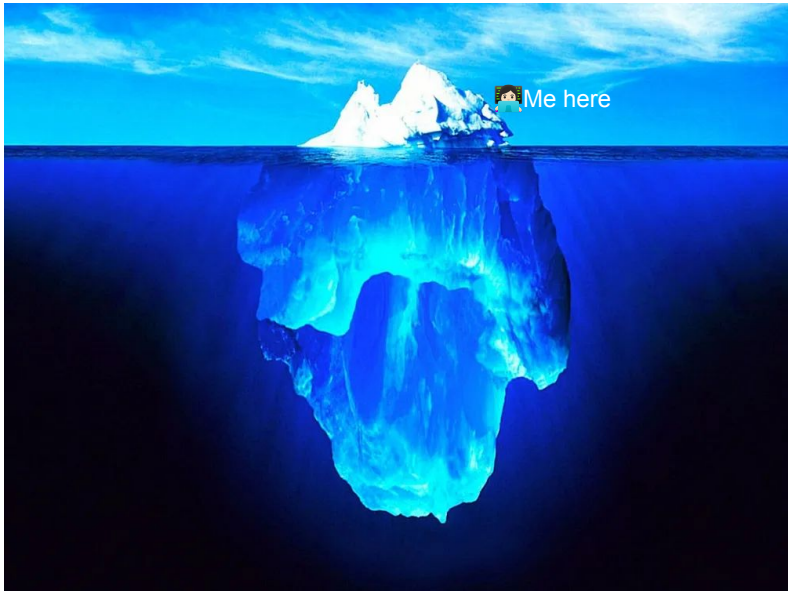


Without Masking, With Masking(128), and Ground Truth

Conclusion

Effect of Attention Map

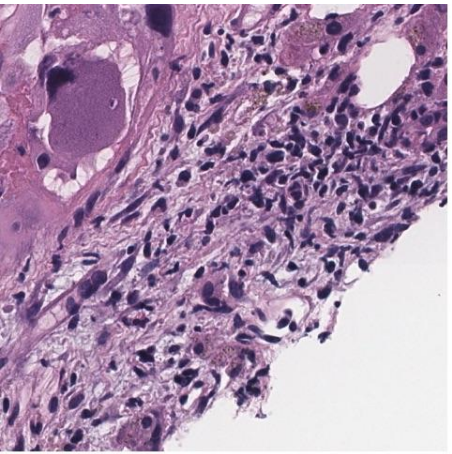
Also, a lot more things we can do with Transformer model

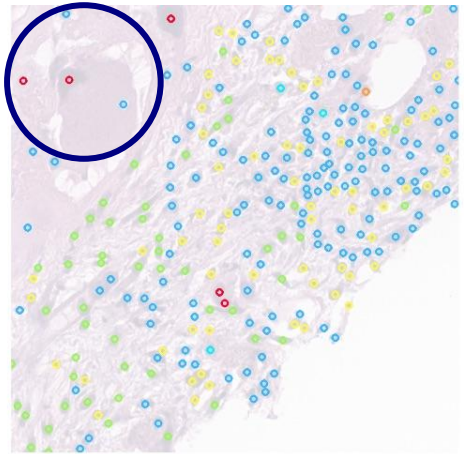


Further studies

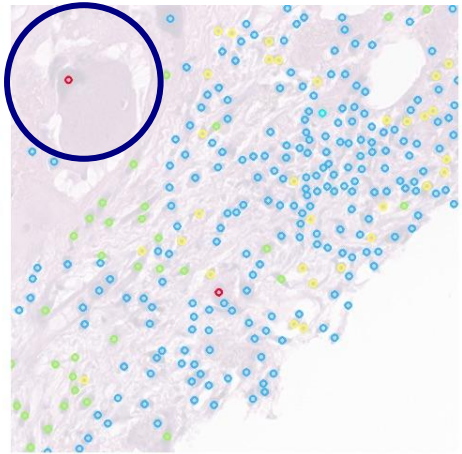
Utilize Padding

Use paddings to retrieve False Negatives of original model

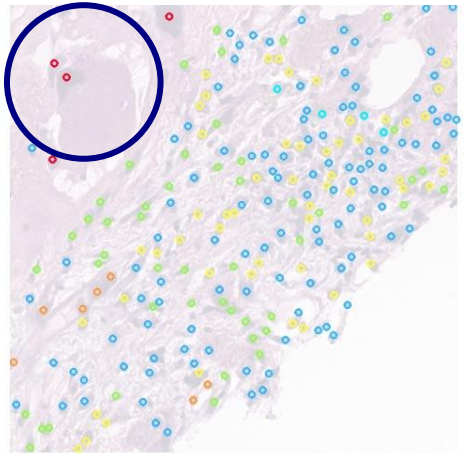




[Orig Pred]



[New Pred]

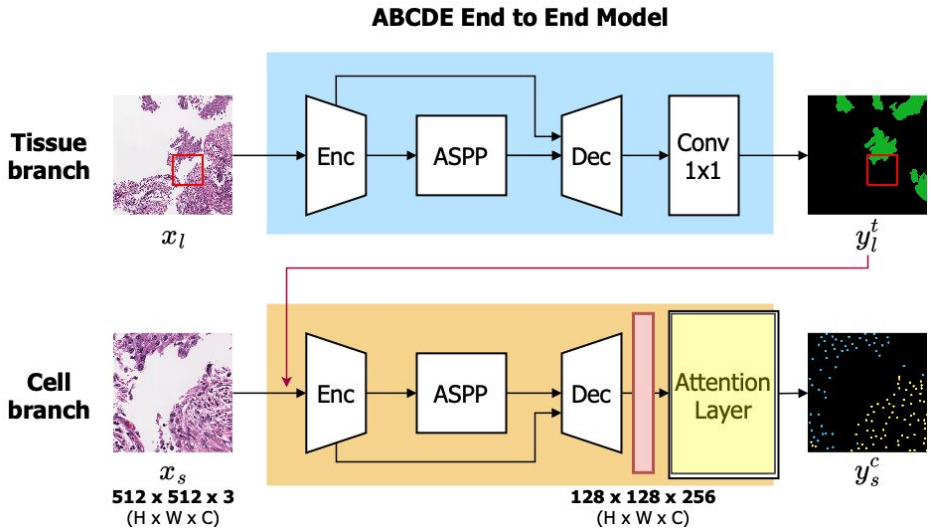


[Label]

Further studies

Train End to End

Instead of using feature map, start from original image and add transformer layer



It was my first time to actually use Transformer

I went through lots of issues and made a lot of bugs🪲 (still fixed some until this week)

Everything was process of learning📚

Thanks a lot to my mentor JWoong💫 to help me get through all those steps

Special thanks to Seonwook, Jinhee and Heon for nice comments during All hands meetings

It was a precious time talking and interacting with all other members of ONCO AI team

Thank you for your Attention :)

Attention Based Cell Detection Enhancer

Jeongin Park

Intern, Model-centric AIR 1, Oncology, Lunit

2023.12.26 ~ 2024.02.29

Lunit

Appendix

Attention Based Cell Detection Enhancer

Jeongin Park

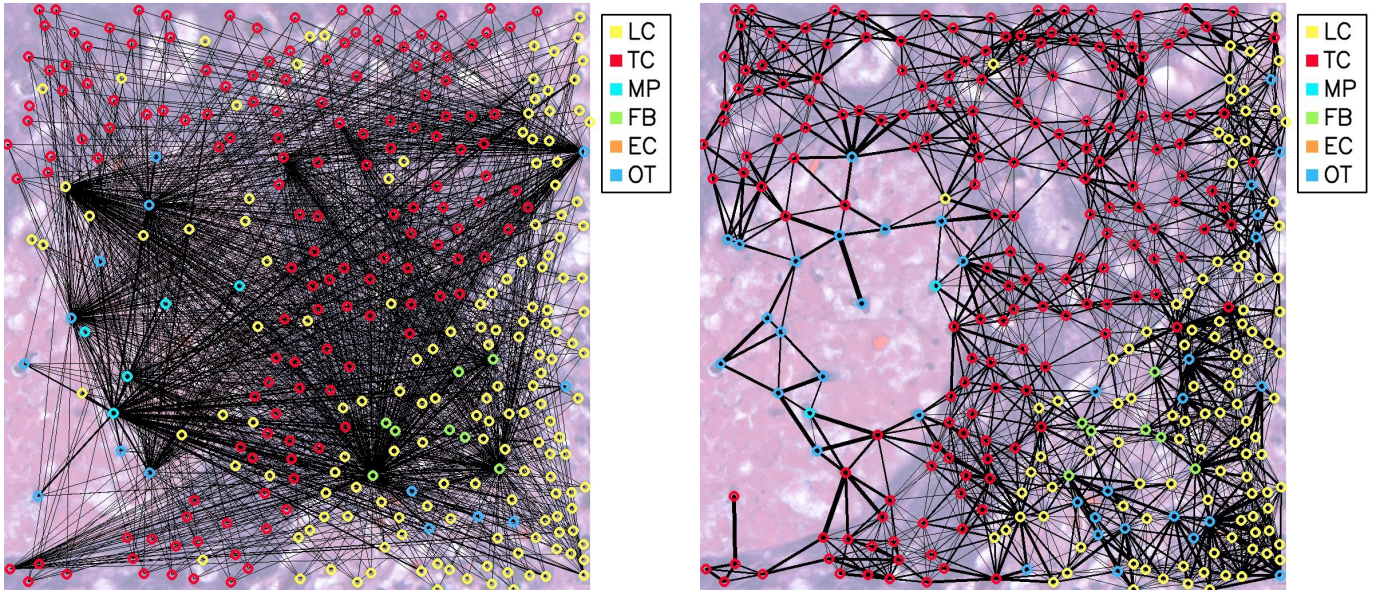
Intern, Model-centric AIR 1, Oncology, Lunit

2023.12.26 ~ 2024.02.29

Lunit

More examples with Attention Map

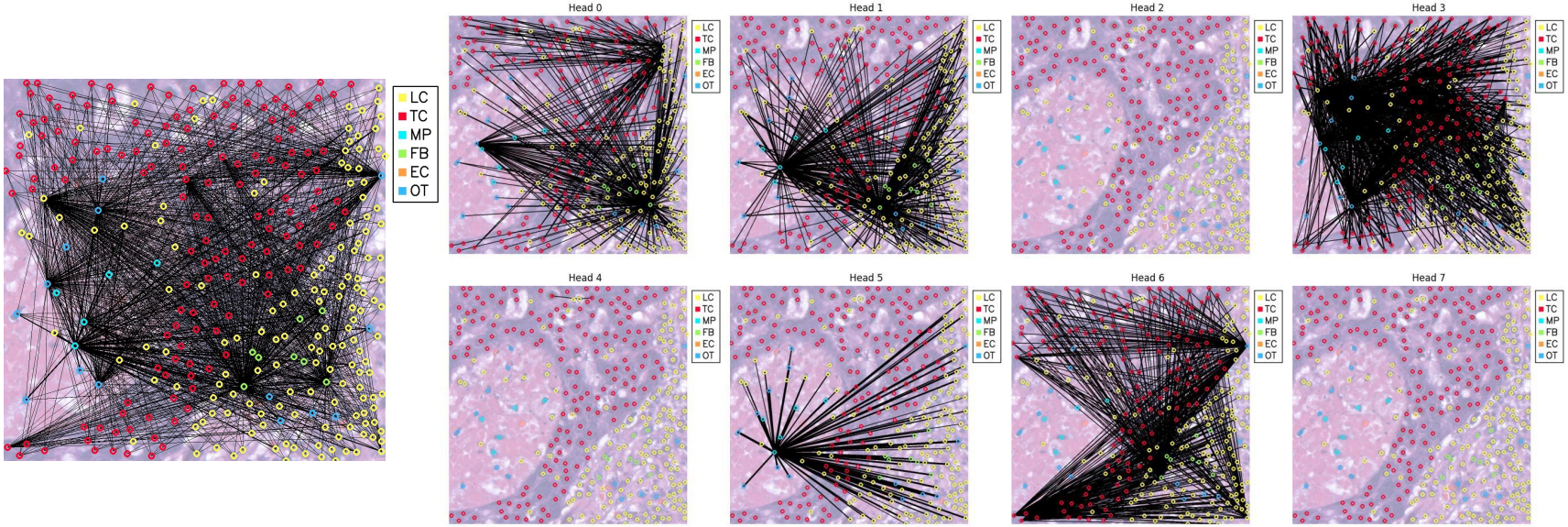
Comparison between no-masking and masking(128)



Average Attention map of each head

More examples with Attention Map

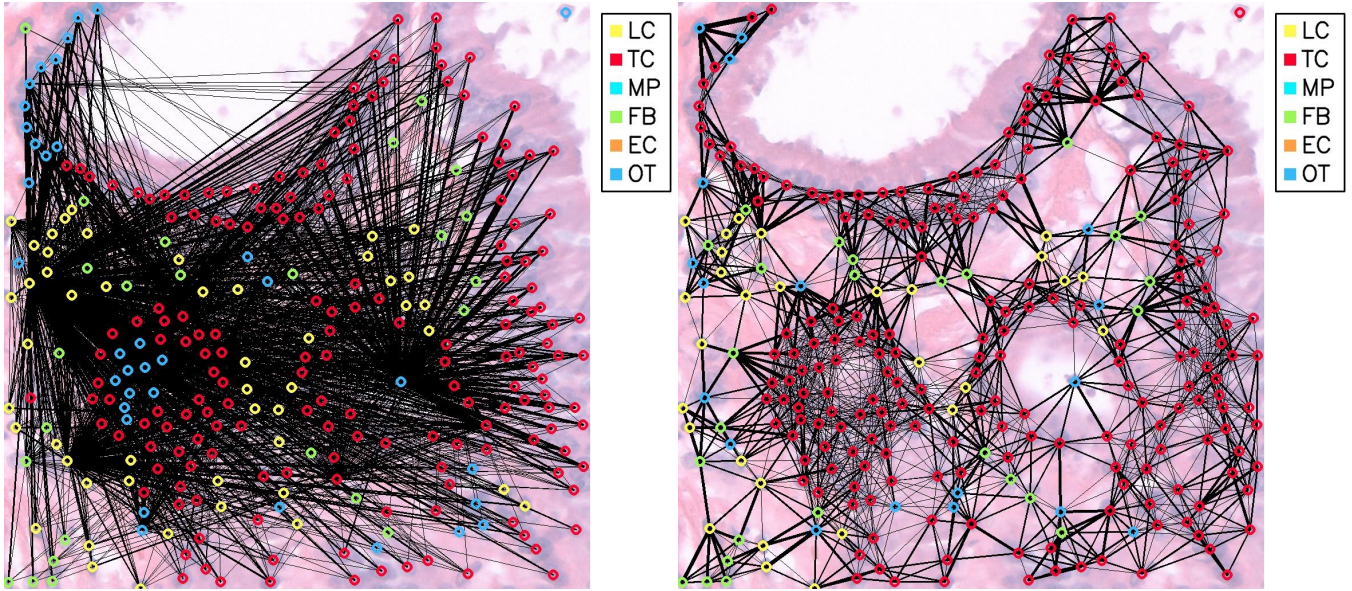
Model without masking



Average attention map, Attention map for each Head

More examples with Attention Map

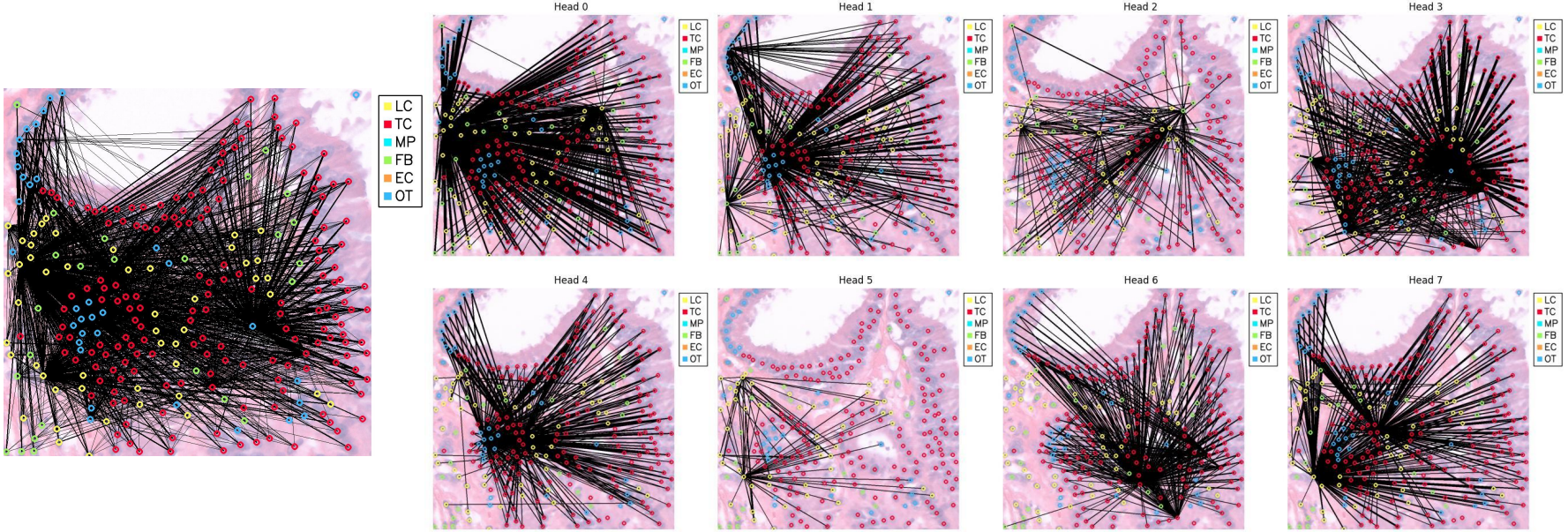
Comparison between no-masking and masking(128)



Average Attention map of each head

More examples with Attention Map

Model without masking



Average attention map, Attention map for each Head