The University Of Queensland AUSTRALIA | CREATE CHANGE

Can AI Learn to See Like Humans?

Frequency-Based Vision for Medical Imaging

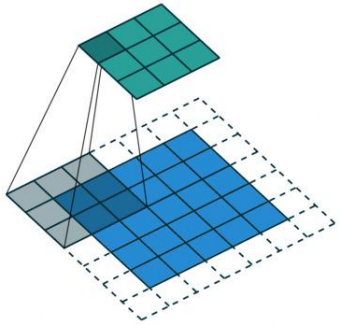
Youssef Hassan, BCompSci/ MCyberSec
Jevi Waugh, MDataSci
Wendi Ma, PhD Candidate
Research Supervisor: Dr. Shekhar “Shakes” Chandra

Acknowledgement of Country

- The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.
- We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.
- We recognise their valuable contributions to Australian and global society.

Image: Digital reproduction of *A guidance through time* by Casey Coolwell and Kyra Mancktelow

Why Rethink Vision Models?

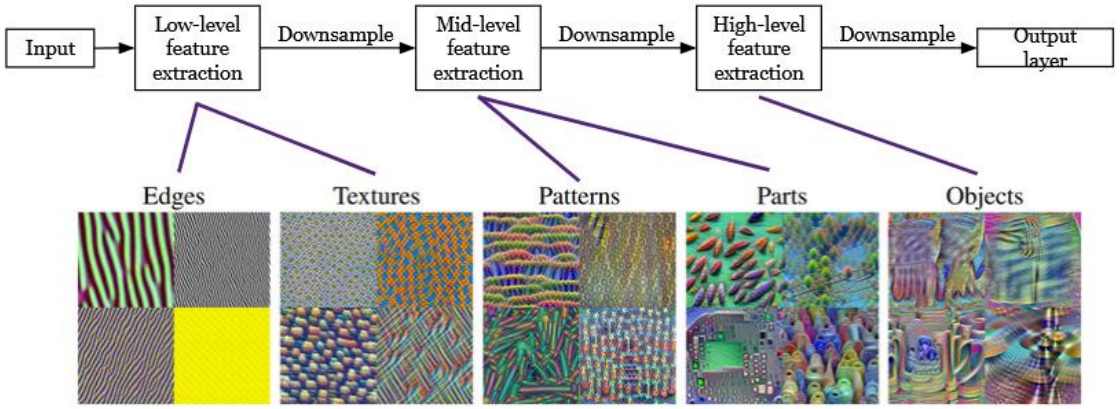


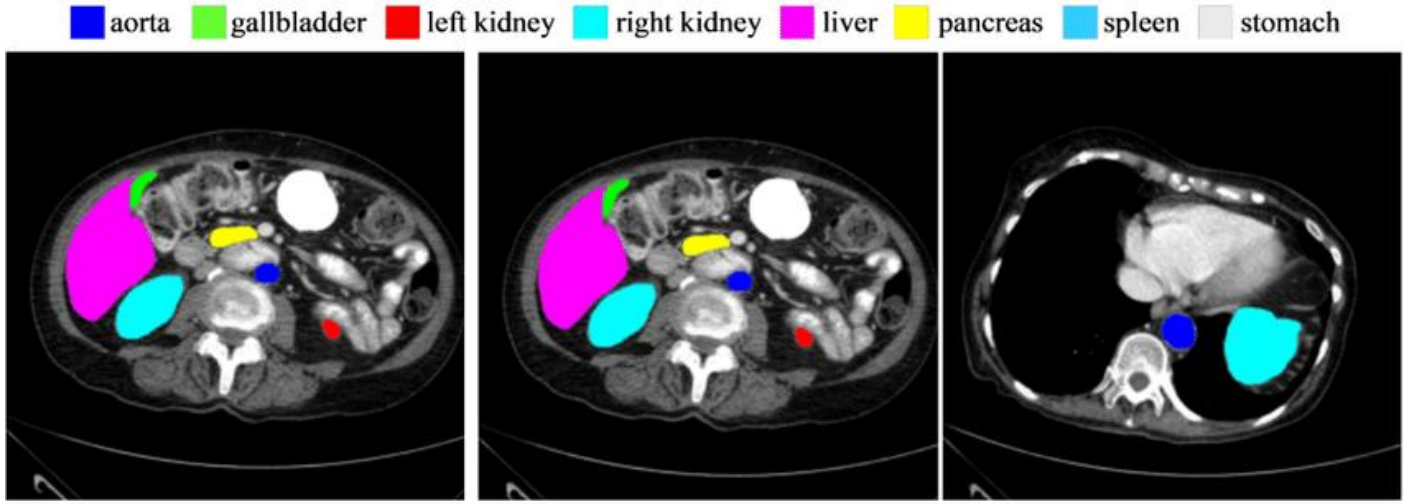
Today's AI sees the world like this.

Powerful. But expensive. And hard to understand.

Segmentation is hard!

Is image-space the only way?



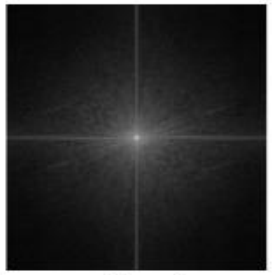


The Frequency Domain

Images are also waves!

Discrete Fourier Transform

$$X[u,v] = \frac{1}{N^2}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} x[m,n]e^{-2\pi i\left(\frac{um}{N}+\frac{vn}{N}\right)}$$

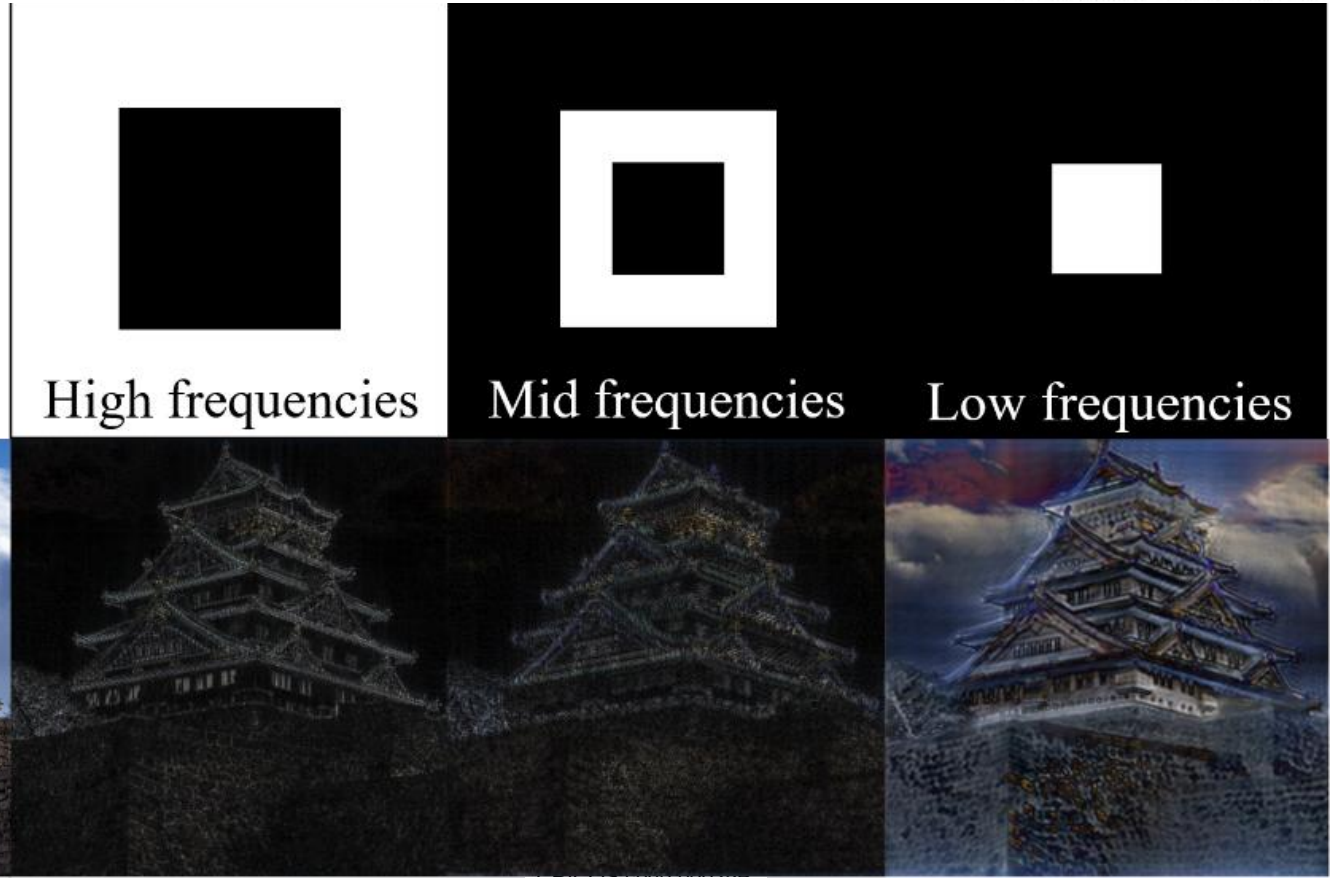


Frequency domain spectrum

Low frequencies correspond to shapes.

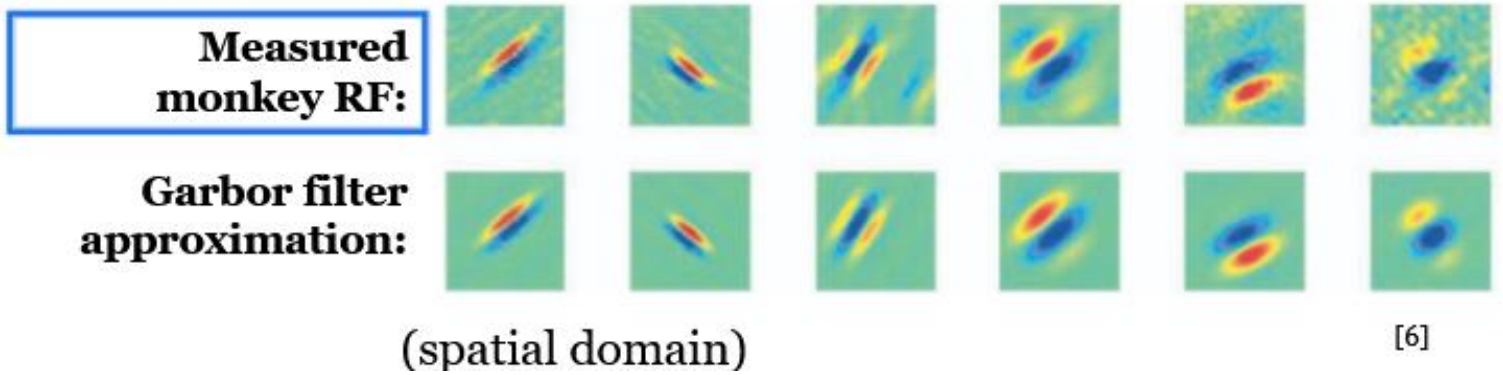
High frequencies correspond to edges.

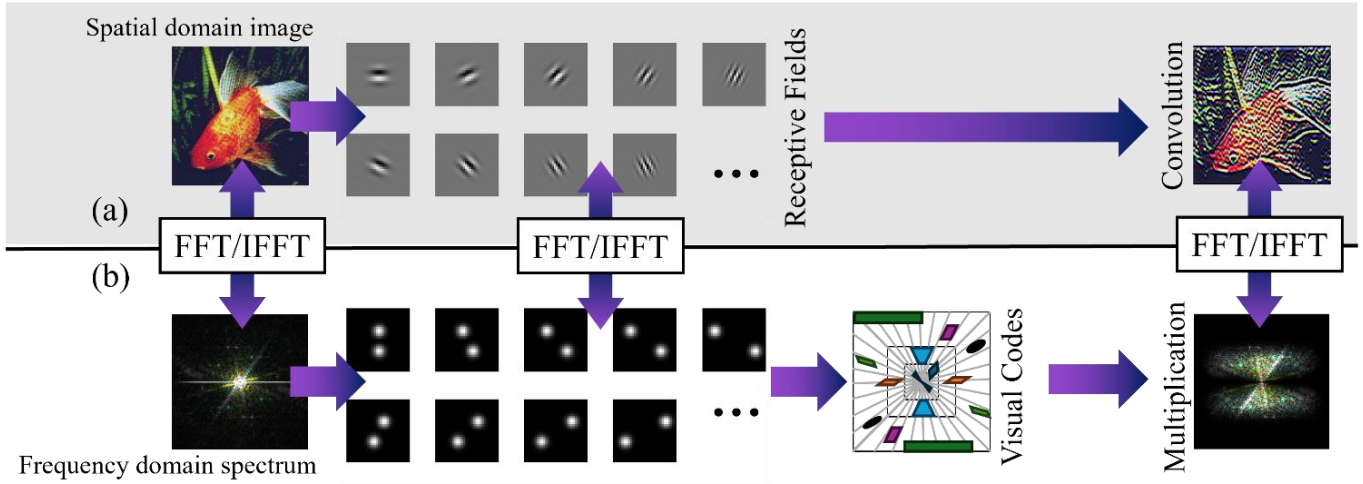
Frequency domain mask

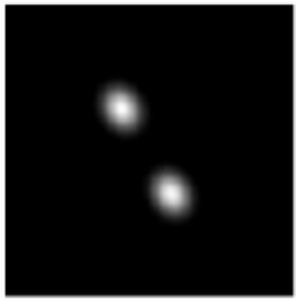


Inspiration from Human Vision

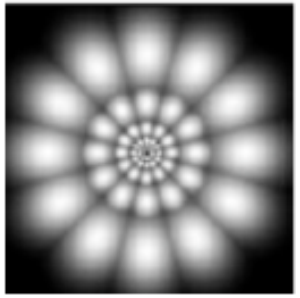
Your brain does not see pixels; it responds to frequencies!







Single Garbor filter (frequency domain)



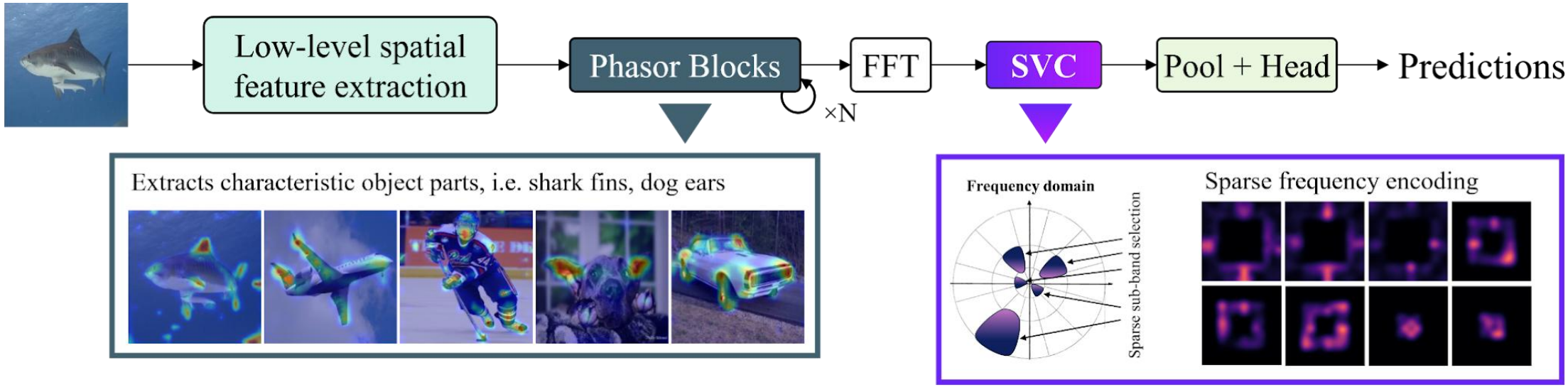
Garbor filter bank (frequency domain)

How PsychoNet Thinks Differently

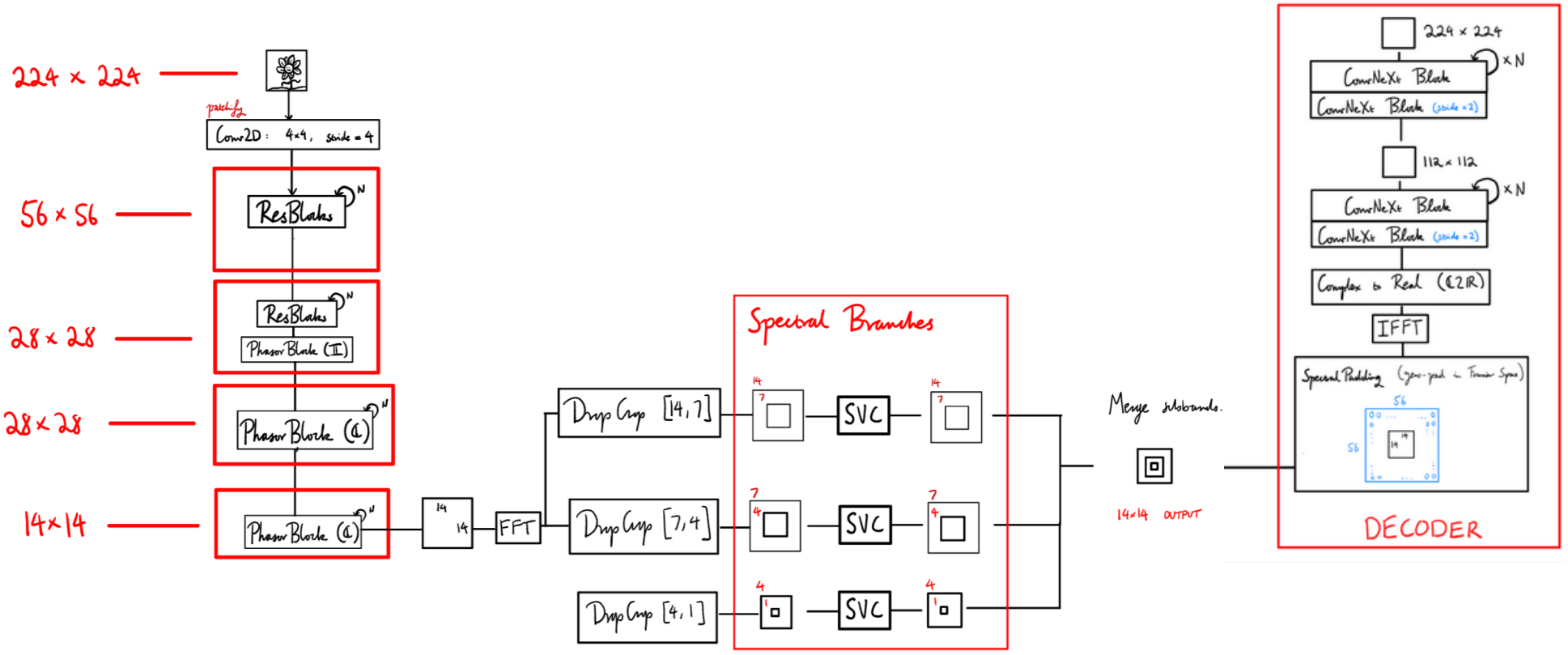
Learning What Frequencies Matter

What It Does

- Converts features into frequency space (FFT)
- Splits into radial bands (low → high)
- Learns sparse frequency codes



The Base PsychoNet Model



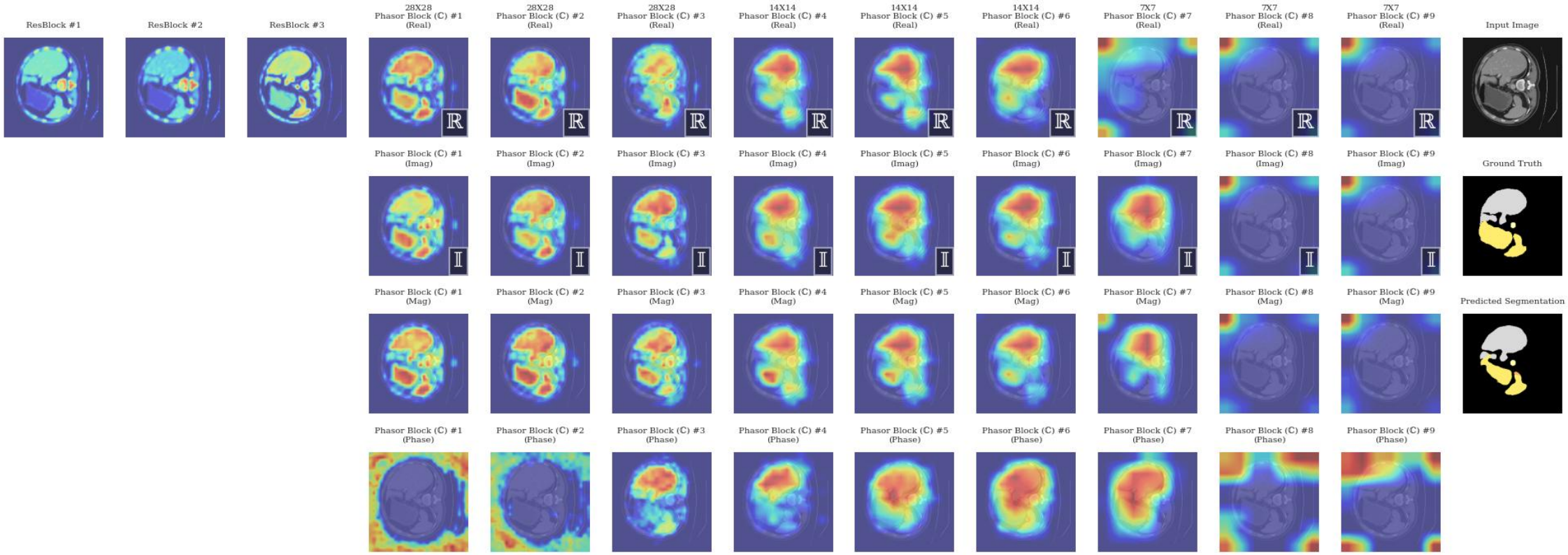
We Tried… A Lot

We joined a project that had been ongoing for a year and was just beginning to work on this new research angle. So, we explored many different pathways.

But one realization changed it all…

Wait… Why Does This Already Look Like a Mask?

14x14 and 7x7 layers were redundant.



Why Were 14×14 and 7×7 Layers Redundant?

Spatial models:

Different layers capture different scales

Require deeper models

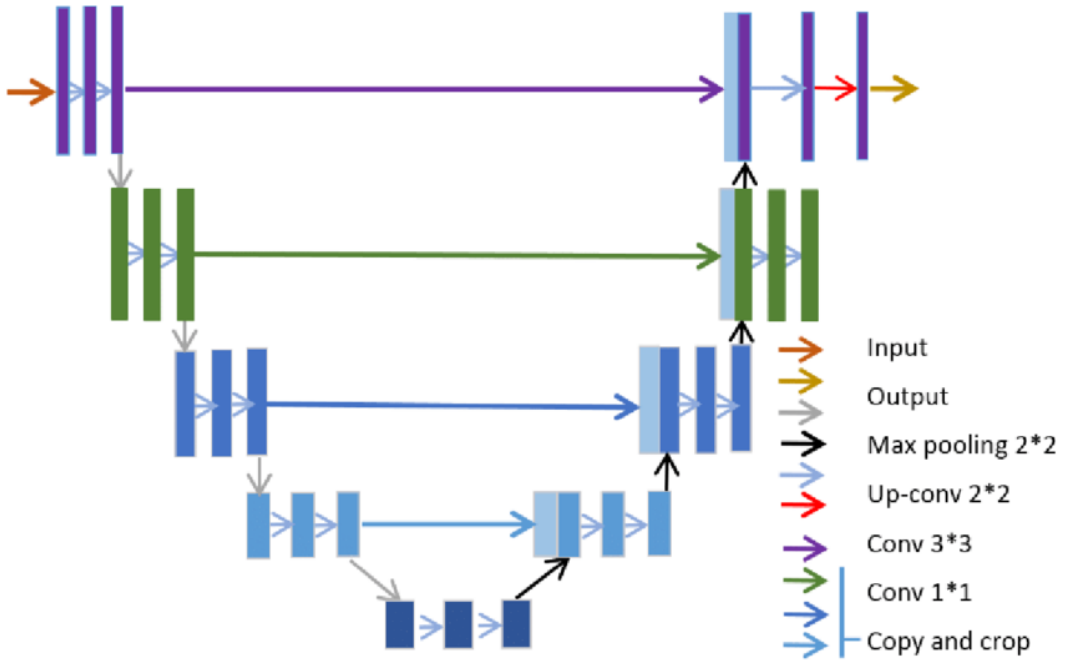
Frequency models:

Capture image-wide features

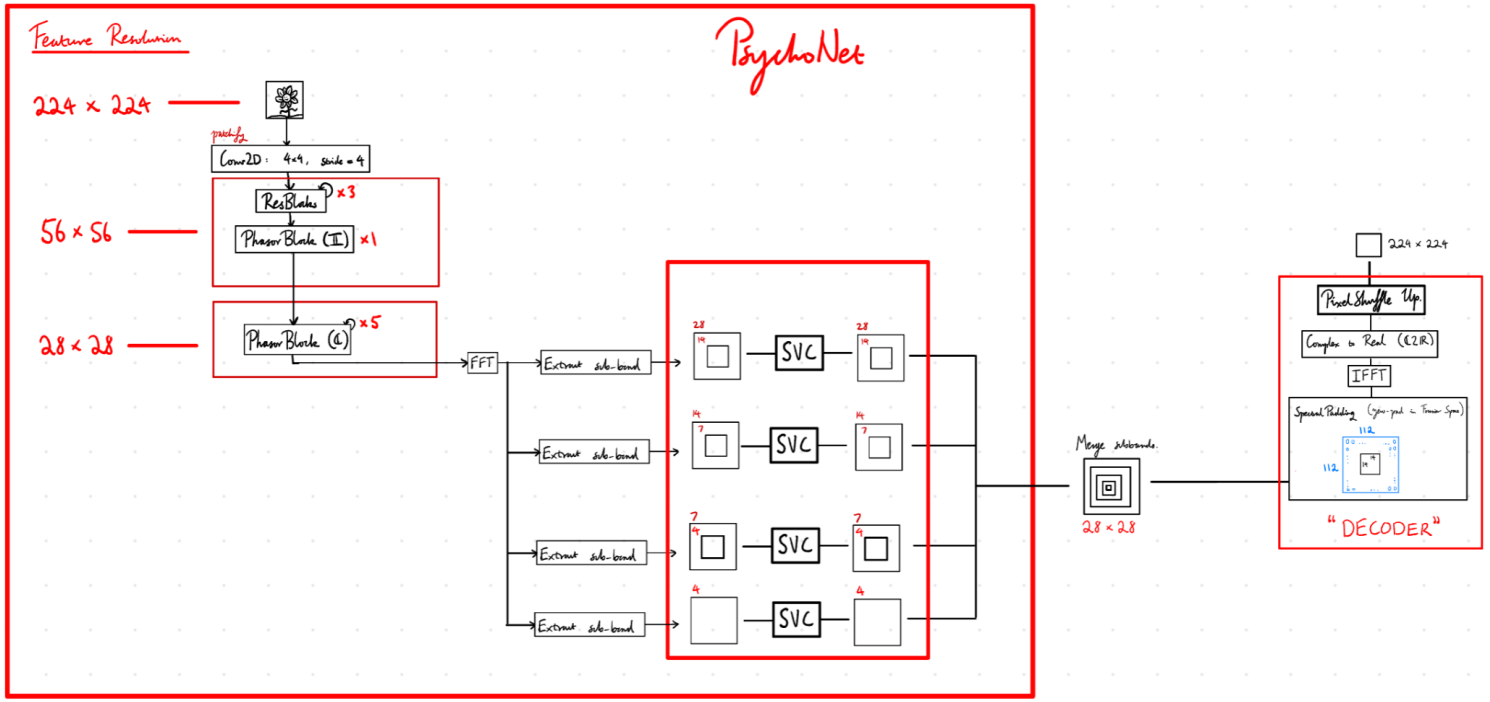
Can be much shallower

Therefore:

We don’t need larger models.



So We Deleted Half the Model.



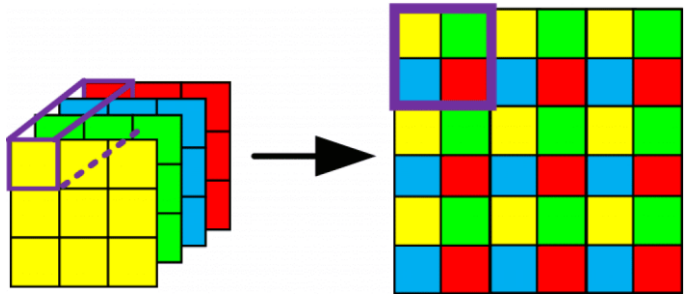
Removing Deeper Layers

- Removed 14×14 and 7×7 stages
- Reduced parameters
- Maintained competitive performance

What we Learned

Key Findings

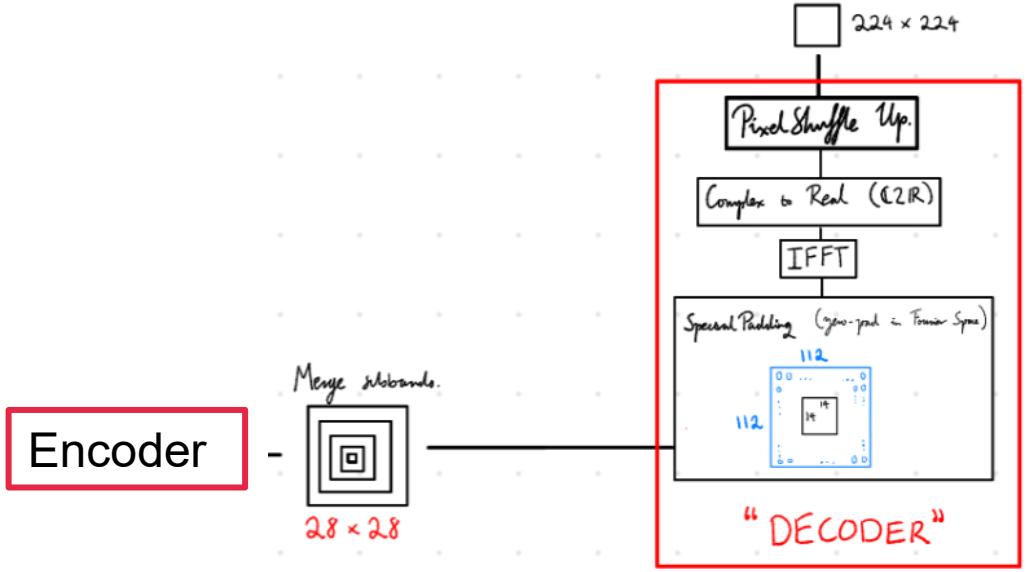
- Frequency representations can support segmentation
- Decoder complexity is less important
- Simpler models can be competitive



PixelShuffle Instead of Convolutions

- Simplified upsampling
- Achieved near-UNet performance
- Segmentation without a “proper” decoder.

This led to the pinnacle of our project…



The Fourier Shuffle

Pixel-Shuffle in Fourier Space

Jevi Waugh

February 2026

1 Introduction

We would like an algorithm for Pixel Shuffle within the frequency domain. Below is the standard 2D Discrete Fourier transform.

$$X[K_y, K_x] = \sum_{m=0}^{H-1} \sum_{n=0}^{W-1} x[m,n] \cdot e^{-2\pi i(\frac{K_y m}{H} + \frac{K_x n}{W})} \quad (1)$$

We would like an expression for

$$DFT(PixelShuffle(x)) \quad (2)$$

$$x \in \mathbb{R}^{(C_{out} \cdot r^2) \times h \times w} \quad (3)$$

X is the spatial data in the shape of [B, C, H, W]. Equation (3) fixes one batch for the sake of the derivation.

$$\# \; c_{in'} = C_{out} \cdot r^2 \quad (4)$$

is the number of channels that goes in the pixel shuffle operation while c_{out} is the number of channels produced after which is defined as: $c_{out} = C$

Now that we have discussed the input, the following shows the output that we want.

$$y = \text{PixelShuffle}(\text{X}) \in \mathbb{R}^{(C \cdot r^2) \times h \times w} \quad (5)$$

Starting with the definition of the Pixel Shuffle:

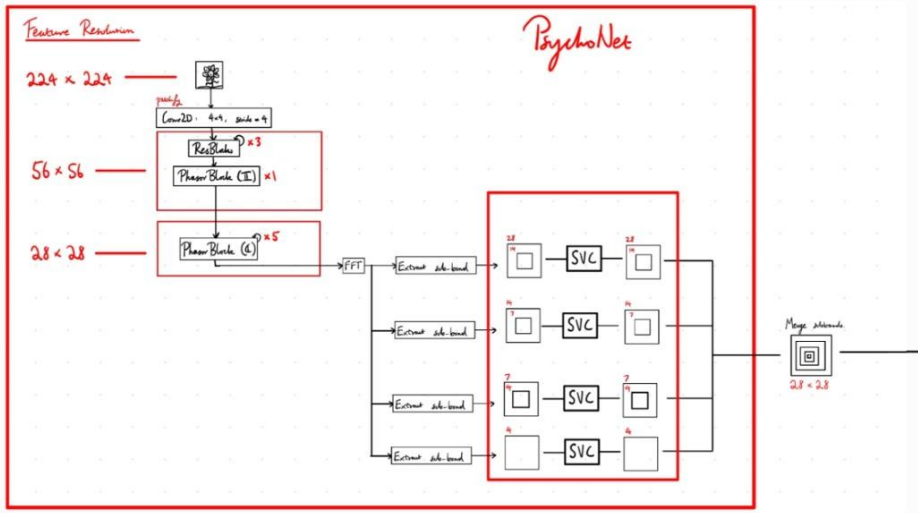
$$y[c_{out}, rh + d_y, rw + d_x] = x[c_{in}, h, w] \quad (6)$$

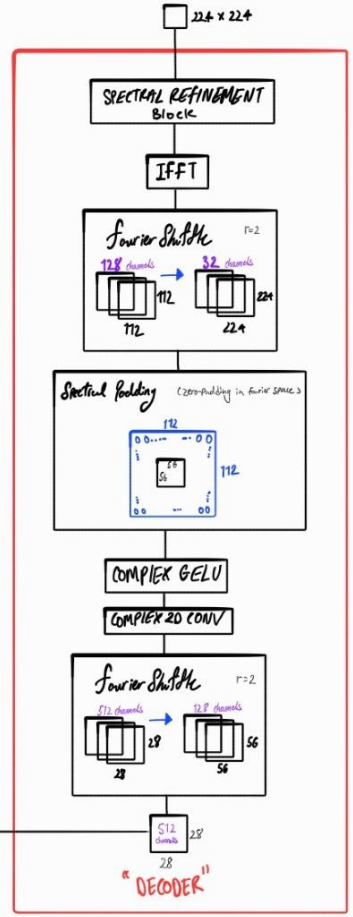
where r is the upscale factor. The following can also be defined:

$$d_x = C_{in} \; mod \; r$$
$$d_y = C_{in} // r \quad (7)$$
$$\text{where } d_x \wedge d_y \in \{0, \cdots, r-1\}$$

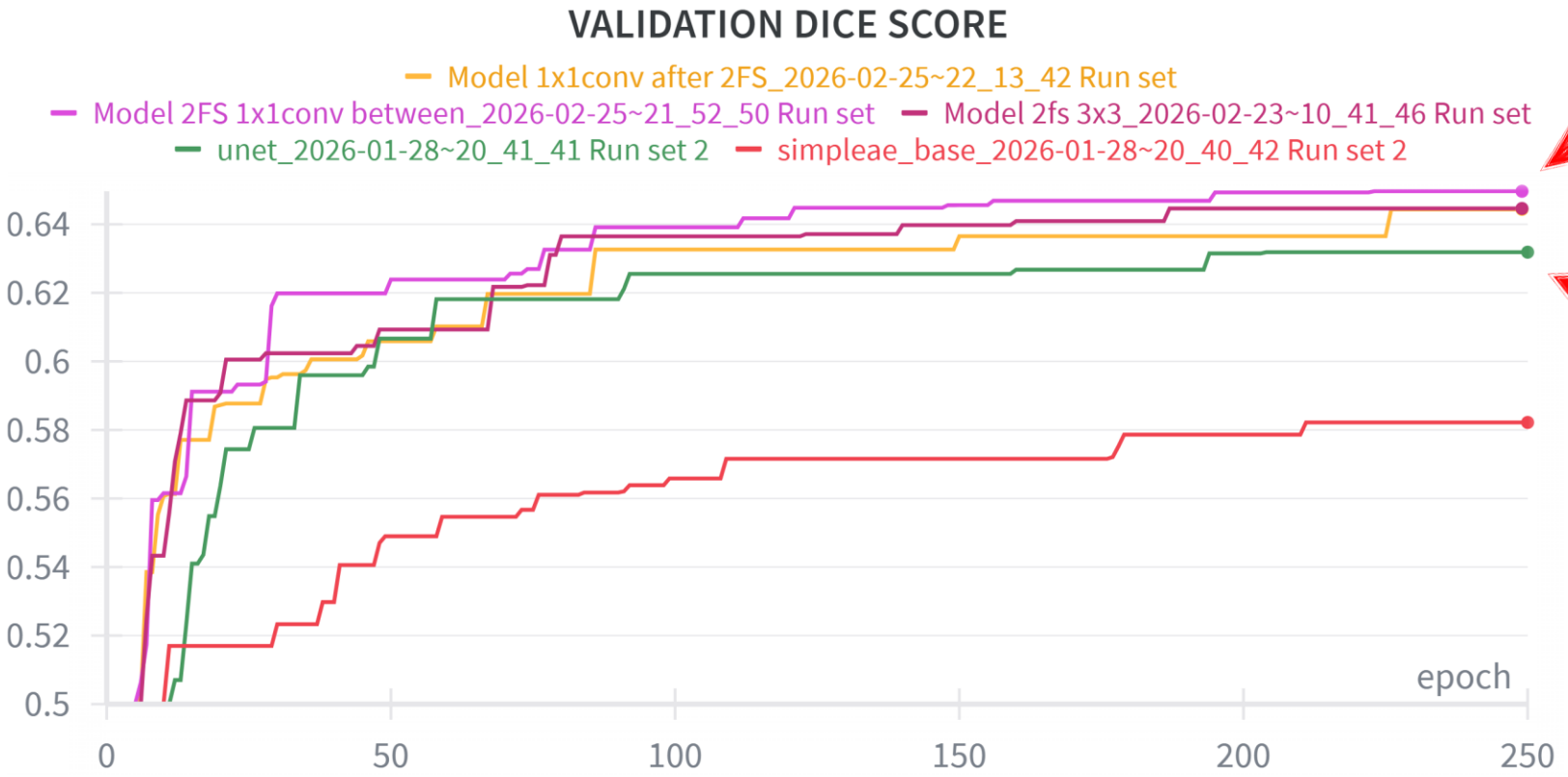
Pages of derivations later…

$$Y[c_{\text{out}}, k_y, k_x] = \sum_{d_y=0}^{r-1} \sum_{d_x=0}^{r-1} X^{\varrho}[c_{\text{out}} r^2 + d_y r + d_x, k_y \; Mod \; h, \; k_x \; Mod \; w] \exp\left(-2\pi i \left(\frac{k_y d_y}{H} + \frac{k_x d_x}{W}\right)\right) \quad (21)$$





Final model: Decoders that beats UNET ON Synapse



$|\Sigma|$

Why This Matters

Current State

Segmentation relies on deep spatial decoders → hard to understand

Possible Future

Shallow frequency-based models

Biologically inspired reasoning

More efficient architectures

Greater interpretability → More Trustworthy AI

Rethinking How AI Sees

Maybe AI doesn't just need deeper networks.
Maybe it needs a different way of seeing.

Youssef Hassan
Jevi Waugh
Supervisor : Dr. Shekhar "Shakes" Chandra
Special thanks to Wendi Ma!
UQ Summer Research