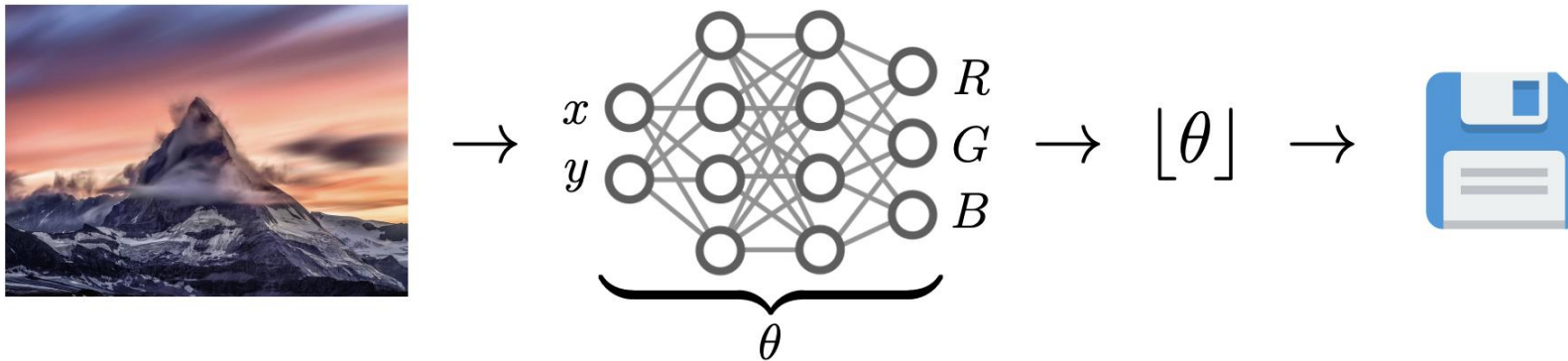


Evaluating Deep Learning Compression Methods for Radio Interferometric Visibility Data in SKAO

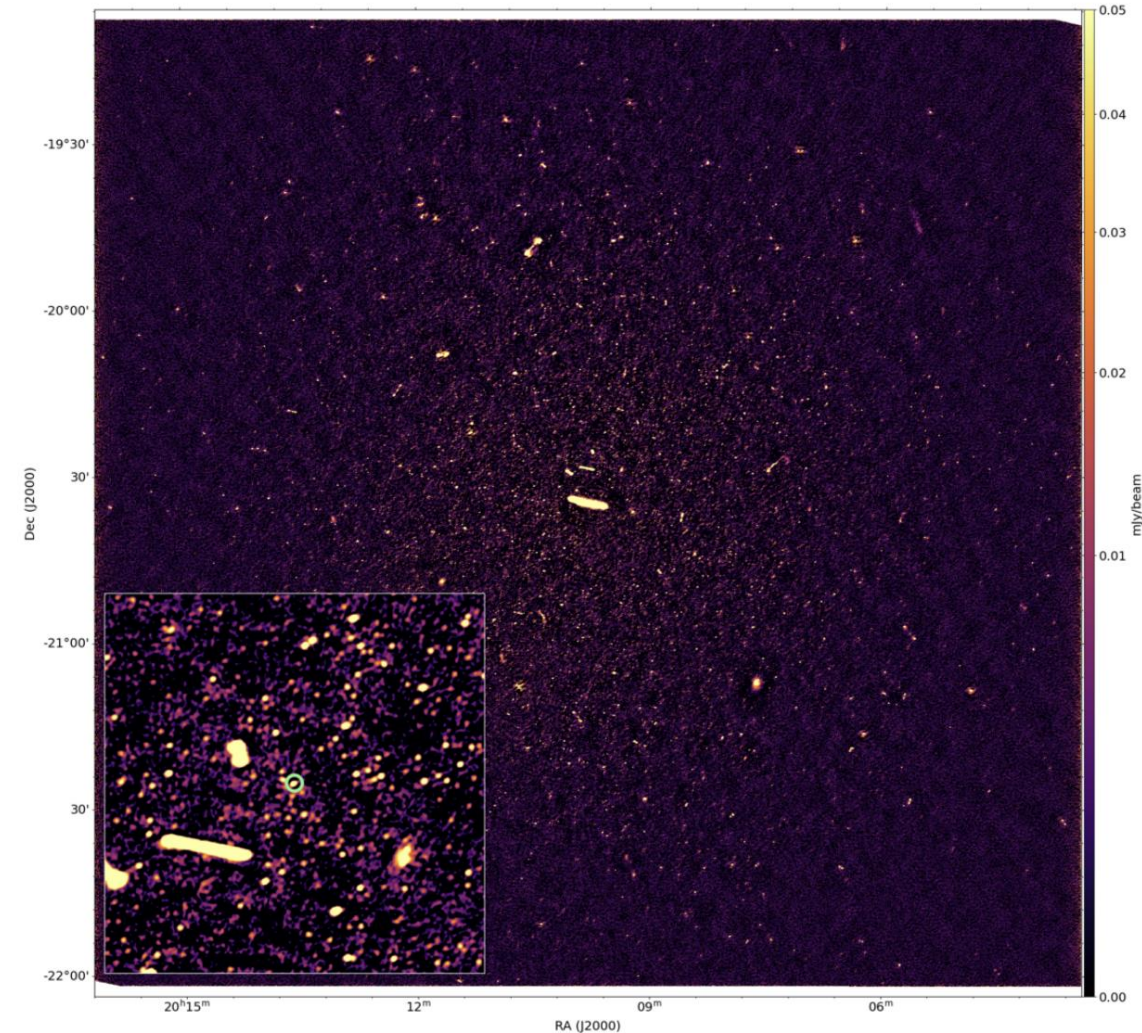
Xiaoran Chen, Adrien Devresse,
Dr Mathieu Salzmann, Prof. Emma Tolley
SDSC, EPFL
SKACH Days 2026

- Implicit Neural Representation (INR) formulation;
 - Represent a signal (image, audio, 3D shape, or here, radio-interferometry visibilities) as a continuous function mapping coordinates to signal values, parameterized by a neural network (typically an MLP)
 - Network weights themselves are the compressed representation, replacing the need to store the raw sampled data
 - Trained by overfitting the network to reconstruct the target signal at its sampled coordinates (e.g. pixel (x, y) , time t , or visibility (u, v, w))
- Applications;
 - Image, audio, and video compression (e.g. COIN, Dupont et al., 2021)
 - 3D shape and scene representation (e.g. NeRF, occupancy/SDF networks)
 - Scientific data compression, e.g. radio-astronomy visibility data (this project, SKA-scale datasets)

- COIN (Dupont et al., 2021), Implicit neural representation model for data compression without entropy coding.
- Represents the signal as a neural network rather than storing explicit values.
- Data values are reconstructed by querying coordinates through **MLP** and **SIREN layers**.



- PARROT data
 - Real MeerKAT visibility data: serendipitous discovery of pulsar PSR J2009-2026 ("PARROT"), from 2020 Saturn–Jupiter conjunction imaging
 - Realistic large-scale test case: genuine calibration artifacts strong scintillation-driven variability
 - Observations: 1 discovery run (2020-12-21, 9.4h) + 7 follow-ups (2021-06 to 2022-02), L-band (856–1712 MHz) & UHF (544–1088 MHz)
 - 4096 channels, 8s integrations, 29–62 dishes -> variable baseline count
 - Pipeline: CARACal (WSClean, CubiCal/QuartiCal, Stimela2), per-subinterval imaging for scintillation

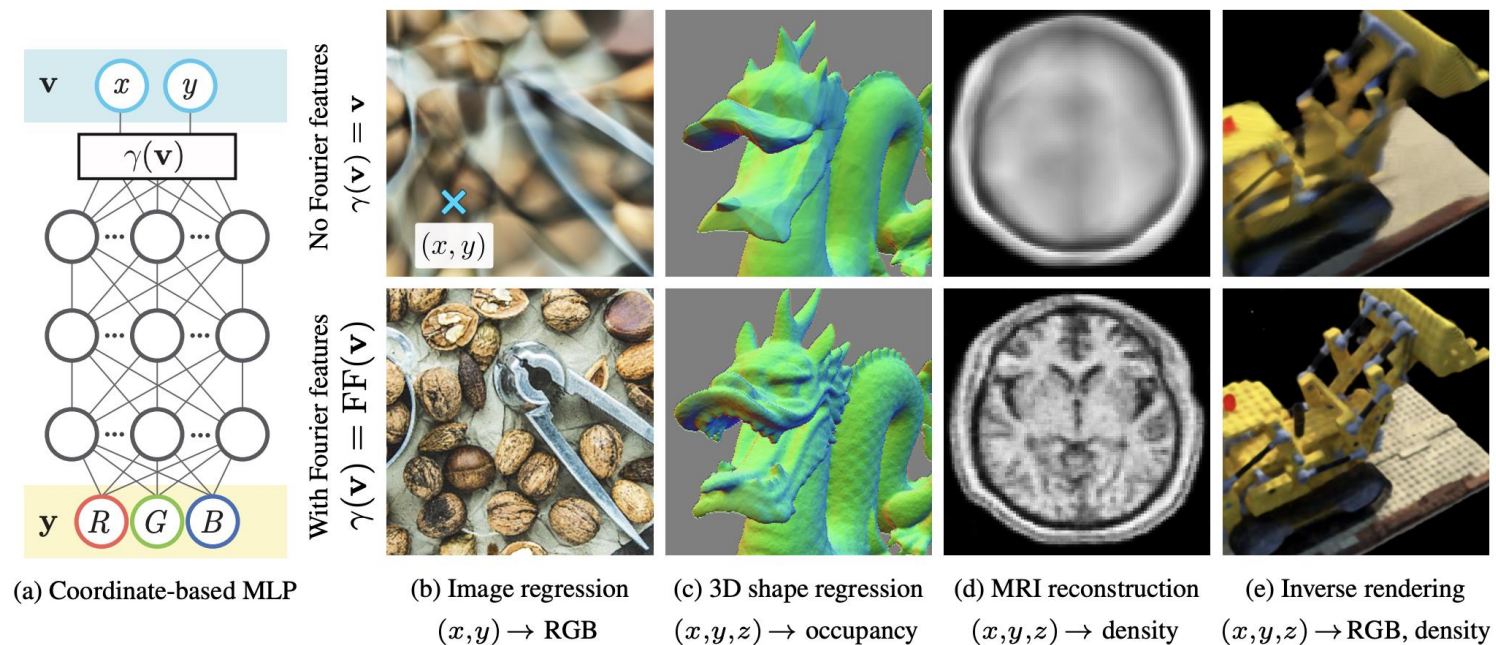


- Baselines for compressing **PARROT** visibility data

Methods	MSE	Mag bias	Phase bias (rad)	Compression rate
DYSCO (Offringa)	0.001	0.026	0.160	34.31%
VAE-based ELIC (He et al.) – per frequency	0.039	0.084	0.561	0.11%
ELIC (He et al.) – per frequency + per antenna	0.011	0.031	0.396	3.00%
INR COIN (Dupont et al.) – per frequency	0.009	0.031	0.779	0.16%
COIN (Dupont et al.) – per frequency + per antenna	0.005	0.064	0.575	8.76%

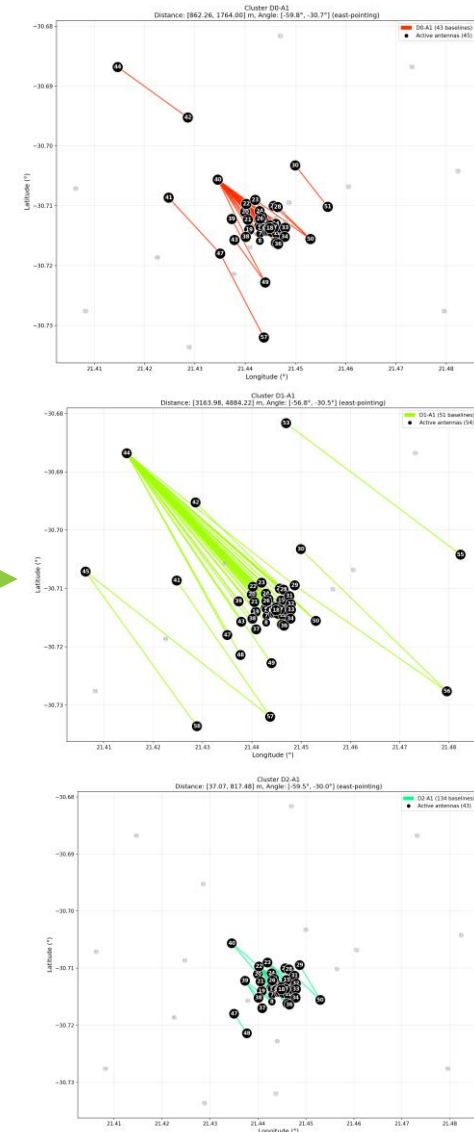
- Natural images: raw coordinates, sufficient to cover high frequency information
 - Pixel grids are smooth and largely band-limited, so an MLP fed raw (x, y) coordinates can already fit most of the target spectrum
 - Positional/Fourier encoding can still speed up convergence, but is not strictly required for good reconstruction quality
- Visibility data: raw coordinates less representative, Fourier features to enhance high frequency information
 - Visibilities are already samples of the sky in Fourier (u, v, w) space, so they are highly oscillatory and broadband
 - Raw-coordinate MLPs suffer from spectral bias and underfit this high-frequency content

- Fourier feature mapping $\gamma(\mathbf{v}) = [\cos(2\pi B\mathbf{v}), \sin(2\pi B\mathbf{v})]$ lifts coordinates into a higher-dimensional space before the MLP, enabling it to learn high-frequency functions (Tancik et al., 2020)
- Frequency scale (bandwidth B) is a key hyperparameter, tuned to match the frequency content of the visibility data



- SIREN (Sitzmann et al. 2020): periodic (sine) activation function
 - $\phi_i(x) = \sin(W_i x + b_i)$ used as the activation at every layer, in place of ReLU/tanh
 - Periodicity lets the network represent fine spatial/temporal detail and model high-order derivatives accurately, since sine is infinitely differentiable
 - Needs a principled weight initialization, uniform in $[-\sqrt{(6/\text{fan_in})}, \sqrt{(6/\text{fan_in})}]$; activation statistics stay stable across layers and gradients don't vanish/explode
 - First layer weights are scaled by a frequency factor ω_0 (typically 30) so the network can span a wide range of spatial frequencies from the start
 - Well suited to implicit neural representations: images, audio, video, and solving PDEs/boundary-value problems via accurate derivatives

- Visibility in freq, baselines, time, u,v,w, infeasible to feed the full dataset to learn one model, too large
 - Full datasets span millions to billions of samples (baselines x frequency channels x time steps)
 - A single INR large enough to fit the whole dataset, computationally challenging;
- Split baselines into clusters: by antenna, **by similarity;**
 - Baseline-similarity clustering groups by (u, v, w) proximity,; nearby baselines see similar sky structure;
 - Antenna-based clustering keeps visibilities from the same antenna pair together



- Computational difficulty to learn complex number directly;
- Split into real + imaginary vs. split into mag + phase;
 - Real + imaginary split: two independent real-valued targets, simple but each still mixes amplitude- and phase-driven variation
 - Magnitude + phase split: separates the smoother amplitude envelope from the phase, which can wrap and needs a periodic (sine/cosine) output representation
- Modelling magnitude and phase separately: each has their own suitable parameters;
 - Magnitude is smoother / lower-frequency, so a smaller network and fewer Fourier features suffice
 - Phase tends to oscillate at higher frequency, benefiting from a larger frequency bandwidth B , more Fourier features, or higher SIREN ω_0
 - Using separate networks (or output heads) for magnitude and phase lets each match its own frequency content and capacity

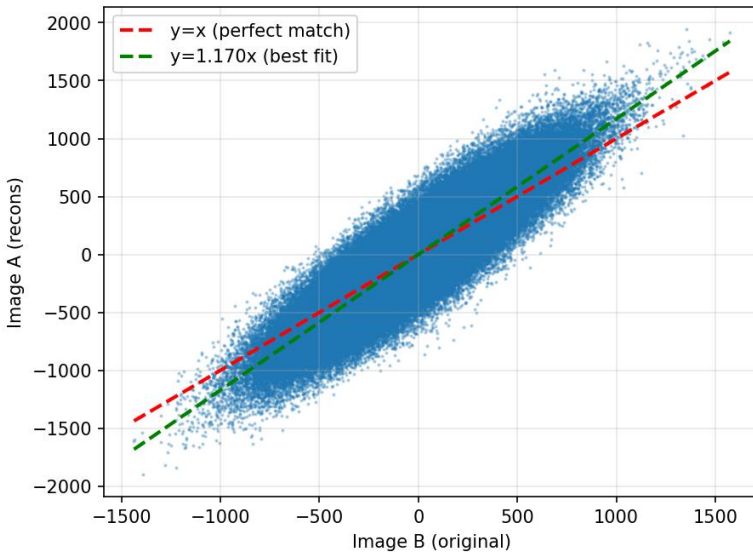
- SIREN layers sensitive to w initialization;
 - Because $\sin(\cdot)$ is periodic, poorly scaled weights push pre-activations outside the range the network can use to approximate the target, causing slow or unstable convergence
 - Small initialization mismatches compound across depth, amplifying layer to layer
- Applying initialization like in WINNER (Chandravamsi et al, 2026) for SIREN layer – brought slight improvement

$$\mathbf{W}_{jk}^{(l)} \sim \mathcal{U}\left(-\frac{1}{\omega_0} \sqrt{\frac{6}{\text{fan_in}}}, \frac{1}{\omega_0} \sqrt{\frac{6}{\text{fan_in}}}\right) \quad W_{jk}^{(l)} \leftarrow W_{jk}^{(l)} + \eta_{jk}^{(l)}, \quad \eta_{jk}^{(l)} \sim \mathcal{N}\left(0, \frac{s}{\omega_0}\right), \quad s = \begin{cases} s_0, & l = 1, \\ s_1, & l = 2, \\ 0, & l = 3, \dots, L. \end{cases}$$

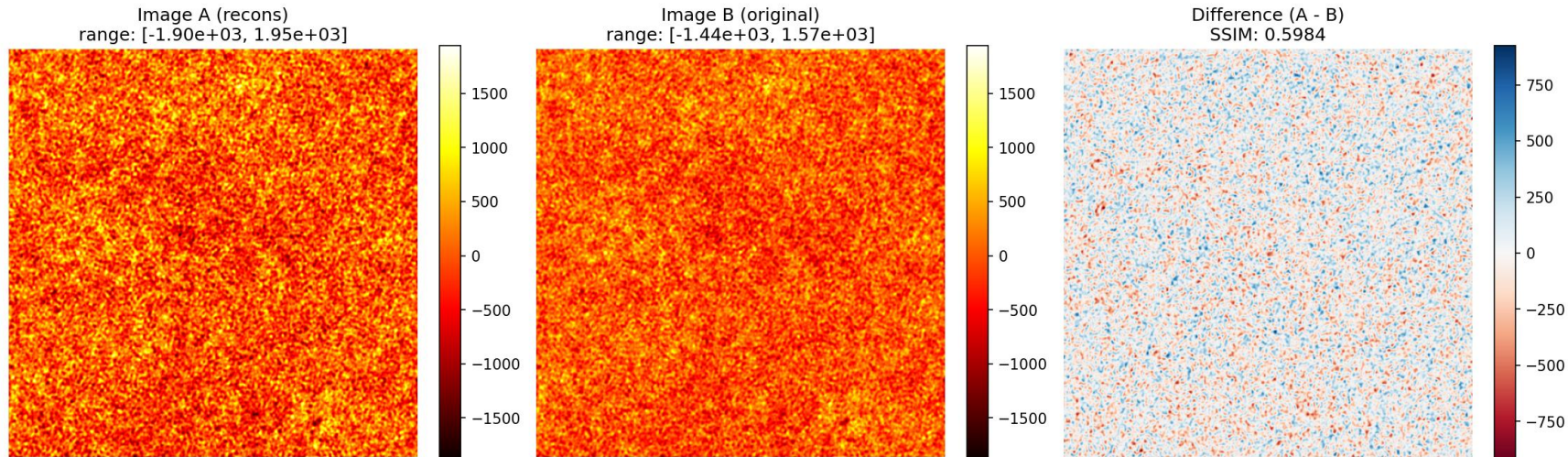
- Sweep search for optimal hyper-parameters S_0 and S_1 for each input image

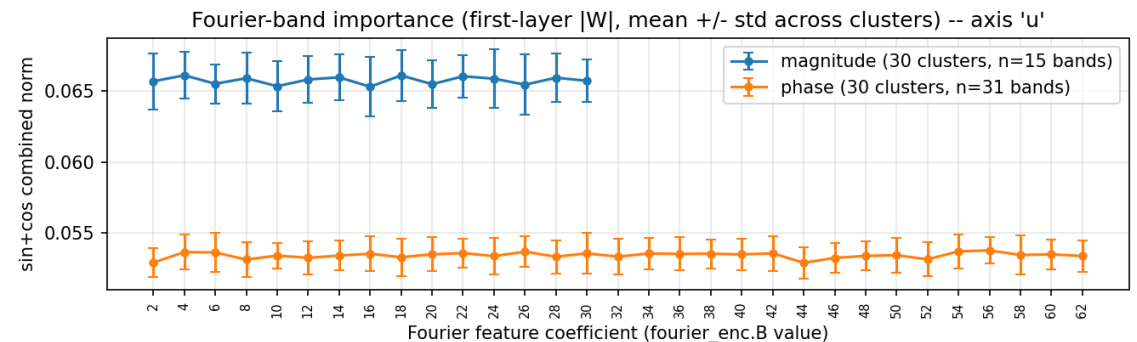
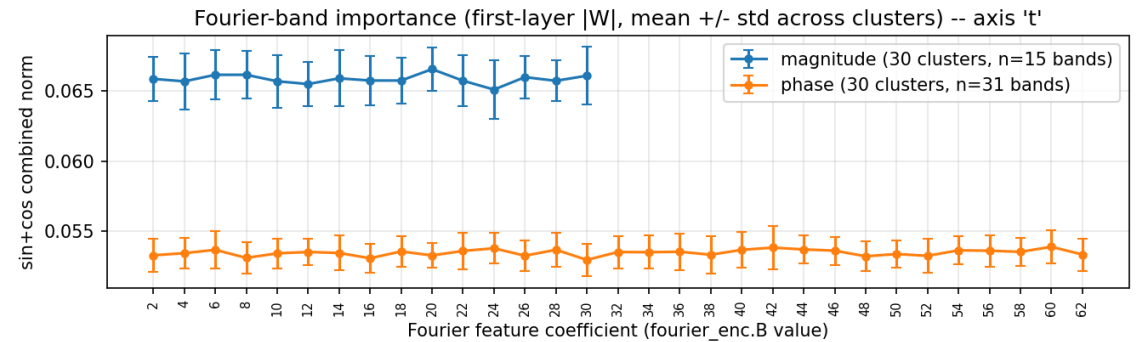
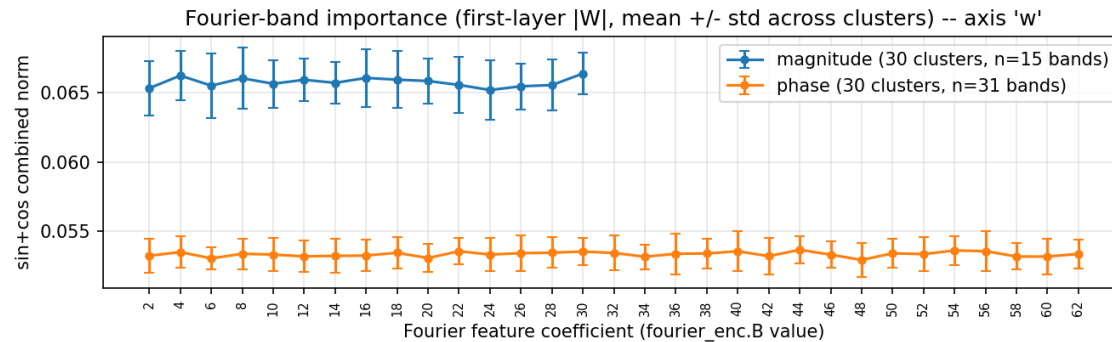
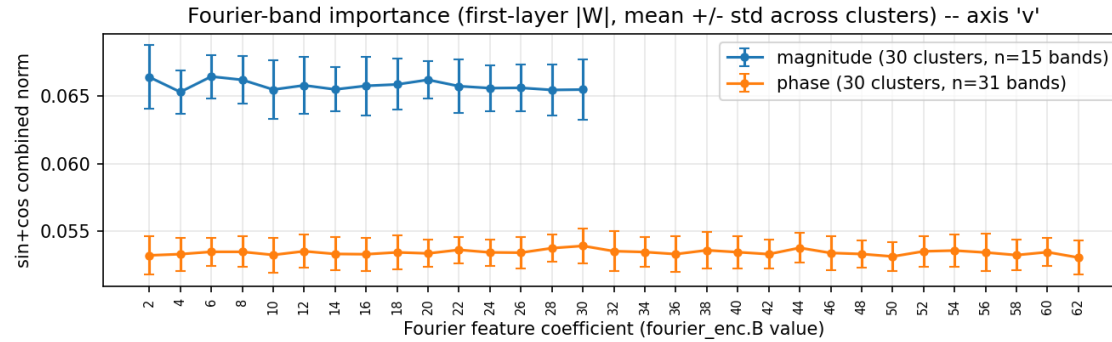
Clustering	Coordinates	# features $\{2k\pi\}$	Layer size	#layers	Mag/phase bias (rad)
antenna	Raw	/	128	10	0.0740 / 0.745
antenna	Encoded	3	128	10	0.0593 / 0.199
Similar baselines	Raw	/	128	10	0.0771 / 0.695
Similar baselines	Encoded	3	128	10	0.0627 / 0.379
Similar baselines	Encoded	16 + 32	128	10	0.0436 / 0.369

Pixel-wise Correlation
Pearson: 0.8902



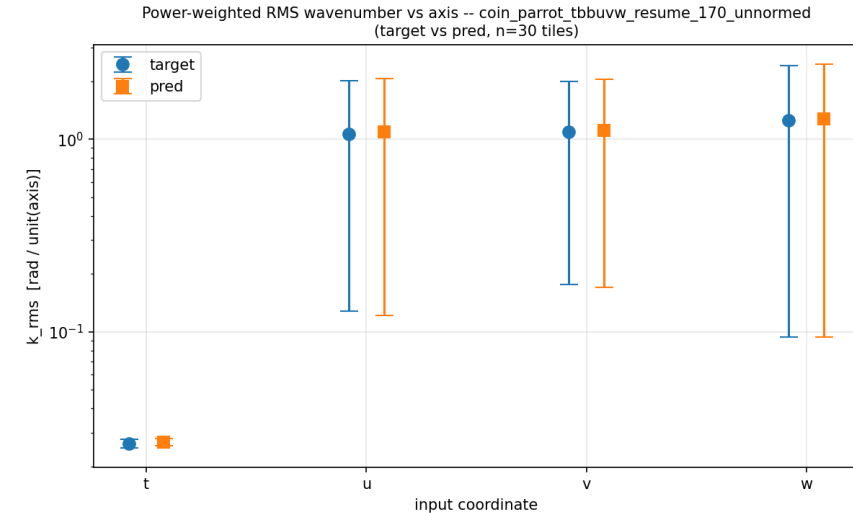
Pearson: 0.8902 | SSIM (orig): 0.5984 | SSIM (norm): 0.8559 | NCC: 0.8902



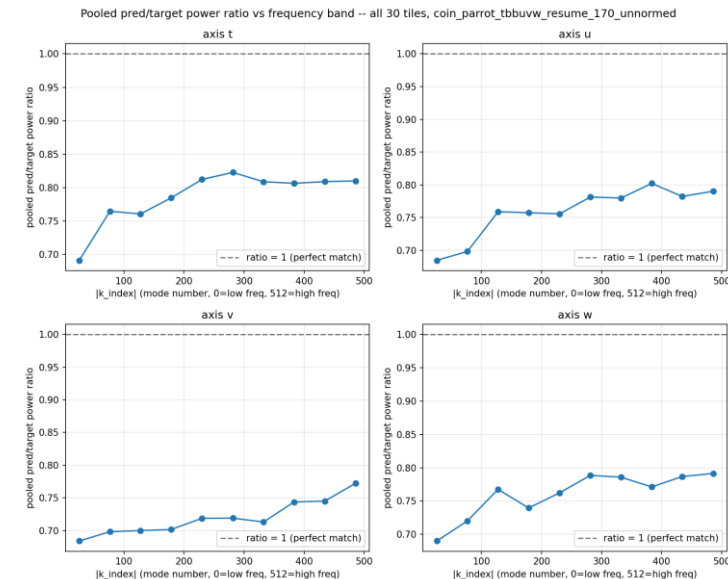


WINNER	Scale*	Layer size	#layers	Mag/phase bias (rad)
without	16 + 32	128	10	0.0436 / 0.369
With ($S_0 = 0, S_1 = 0.5$)	16 + 32	128	10	0.0308 / 0.336

- Improved on the convergence speed;
- Root issue: oscillation captured, but scale (magnitude) is mis-matched



shape



scale

- Worth checking per-antenna clustering without repetition;
- Compressing the whole data is still time-consuming, ~1024 frequency channels;
- Down-stream tasks to assess the feasibility of INR data compression
 - Getting dirty images from all frequency → clean image;
 - Applications..