

Deep-learning in visibility space

Philipp Denzel¹, Thamena Essahaty^{1,2}, Frank-Peter Schilling¹, Elena Gavagnin^{1,3}

¹ Centre for Artificial Intelligence, ZHAW, Winterthur, Switzerland,

² School of Engineering & Computer Science, VUW, Wellington, New Zealand,

³ Institute of Business Information Technology, ZHAW, Winterthur, Switzerland

Swiss SKA Days 2026, Neuchâtel

ZHAW | Centre for Artificial Intelligence

2026-09-01 | Philipp Denzel

Why act in visibility space?

The sampling problem

Van Cittert-Zernike: $\mathcal{V}(u, v) = \mathcal{F}\{A(l, m) I(l, m)\}$

– Correlating at baselines $\rightarrow (u, v)$ sample of Fourier attenuated sky

More (u, v) samples $\rightarrow$ **better** $\hat{I}$

– Finite set of samples: image reconstruction is **ill-posed**

Every method closes the gap with a **prior**

– e.g. **CLEAN** assumes the sky is composed of point sources

– Problematic for faint, **extended**, complex structures

Our goal

– **Generate** missing visibility samples using DL "*at acquisition time*"

– **Denoising and super-resolving** the reconstructed image downstream

– *Learning the prior instead of asserting it*

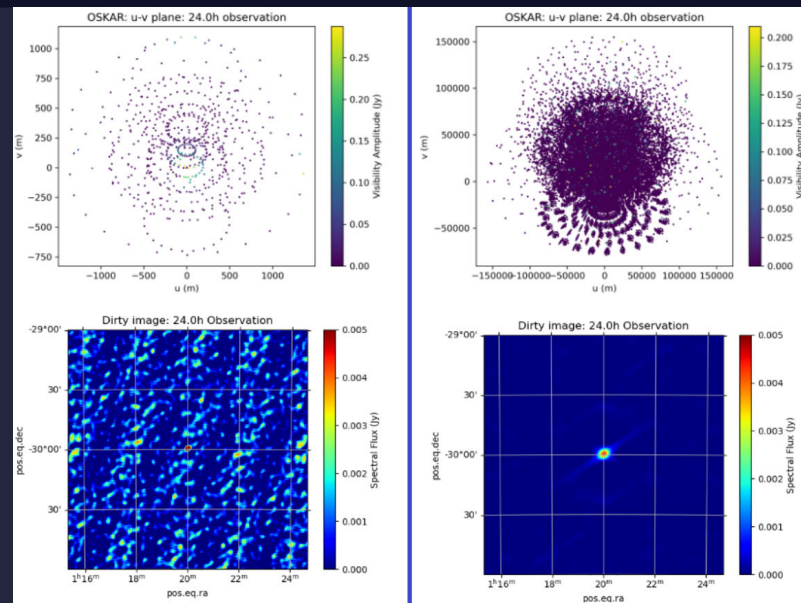
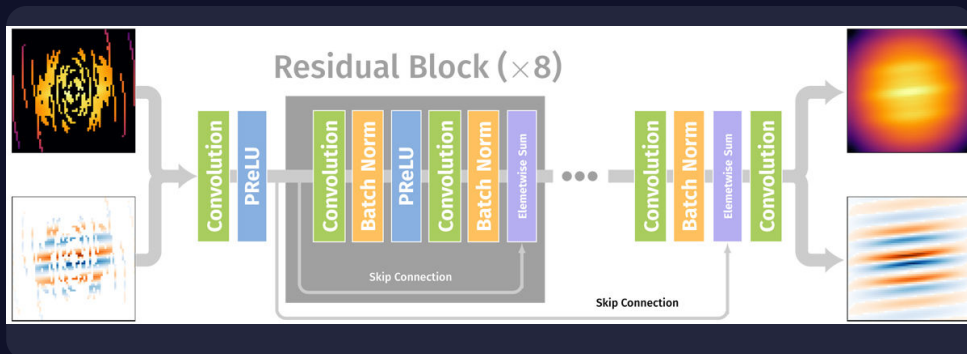


Figure 1: OSKAR, **24 h** observation of the same sky. Left: SKA-Mid-**AA1**, 36 baselines. Right: **AA4**, 19,503 baselines. Top: (u, v) coverage; bottom: the resulting dirty image.

radionets

(Schmidt+22, Fröse+26) — the only close precedent:

- sampled cells in, **complete** (u, v) plane out
- but **one aggregated 2-D plane** per observation — *no time axis*
- trained on a **single array** (VLBA — **10 antennas**, 45 baselines)
- Low number of baselines
- *time axis collapsed* — one plane in, one plane out



radionets approach: masked **amplitude** (top) and **phase** (bottom) of one *finished* (u, v) plane in $\rightarrow$ dense amplitude and phase out, then an inverse FT to the image. **8 residual blocks**, super-resolution CNN.

MSc thesis by Essahaty

(Thamena Essahaty, ZHAW $\times$ VUW Wellington NZ, 2026)

"Inferring Interference: Deep Learning for Radio Telescope Data Enhancement"



Supervised by Philipp Denzel (ZHAW) and Bing Xue (VUW).

Visibilities are large

Visibility tensor

One visibility tensor — SKA-Low:

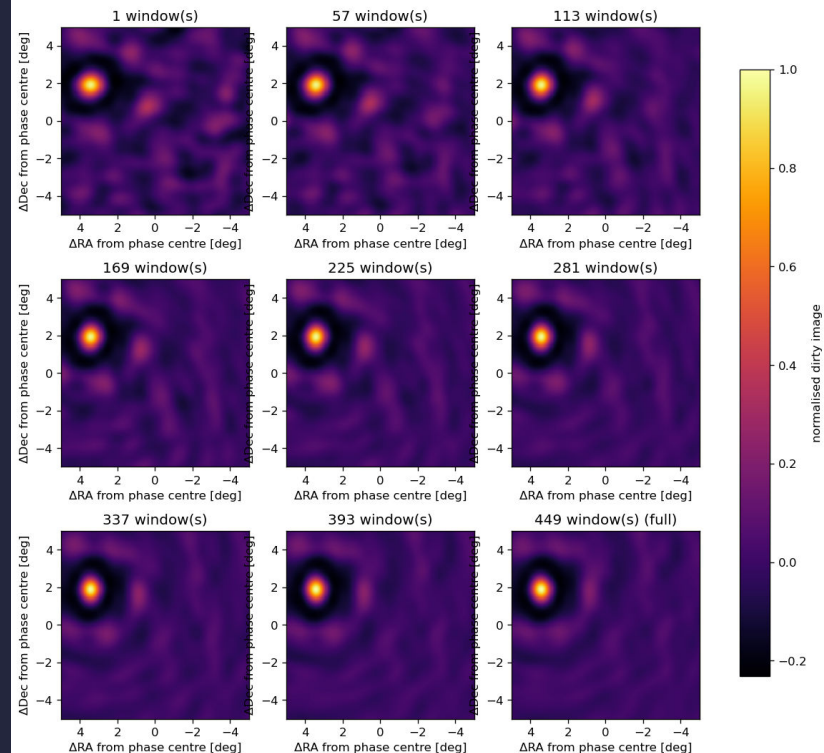
$$V \in \mathbb{C}^{T \times F \times B \times P}$$

- **T** timesteps — @ 1s integrations (3600 for 1h)
- **F** frequency channels — $O(10^2 - 10^5)$
- **B** baselines — $\frac{N(N-1)}{2}$ cross-correlations of N stations/dishes
- **P** polarisations, Stokes I, Q, U, and V

At `complex64`, scaling the baseline axis only ($T=3600$, $F=64$, $P=4$;
window = 16 timesteps, stride 8 → 449 windows):

stage	stations	baselines	file	window	timestep
AA2	64	2,016	13.8 GiB	63 MiB	3.94 MiB
AA*	307	46,971	323 GiB	1.43 GiB	91.7 MiB
AA4	512	130,816	898 GiB	3.99 GiB	256 MiB

dirty image vs uv-coverage — channel 32 (145.0 MHz), Stokes I, window_size=8, stride=8, phase centre (0.000, -27.000)°



An SKA-Low AA2 simulation, 1 → 449 windows (2h): the sidelobes fall away as the Earth turns.

How we posed the task

Three problem formulations

- **Multistep forecasting** — $f_{\theta}(V_{1:t}) = \hat{V}_{t+1:T}$: predict the (u, v) slices not yet measured
- **Fourier-plane super-resolution** — a dense $\hat{V}$ from a sparse V
- **Latent prediction** — compress each slice $V_t \rightarrow z_t$, forecast $\hat{z}_{t+1}$, decode

All three are **3-D** by construction: two (u, v) axes plus time.

Four architectures compared

- **3D U-Net** — convolutional baseline
- **3D patch-based ViT** — spatiotemporal patch embeddings
- **3D slice-based ViT** — one whole (u, v) slice per token, attention across time
- **Attention autoencoder + latent ViT** — forecasting on top of a learned compression

Training

- with a sparsity-aware, flux-weighted loss

Evaluation

- PSNR in the Fourier plane (F)
- PSNR / MSSIM on the reconstructed dirty image (I)

Basic SKA-Low AA2 GLEAM-like simulations

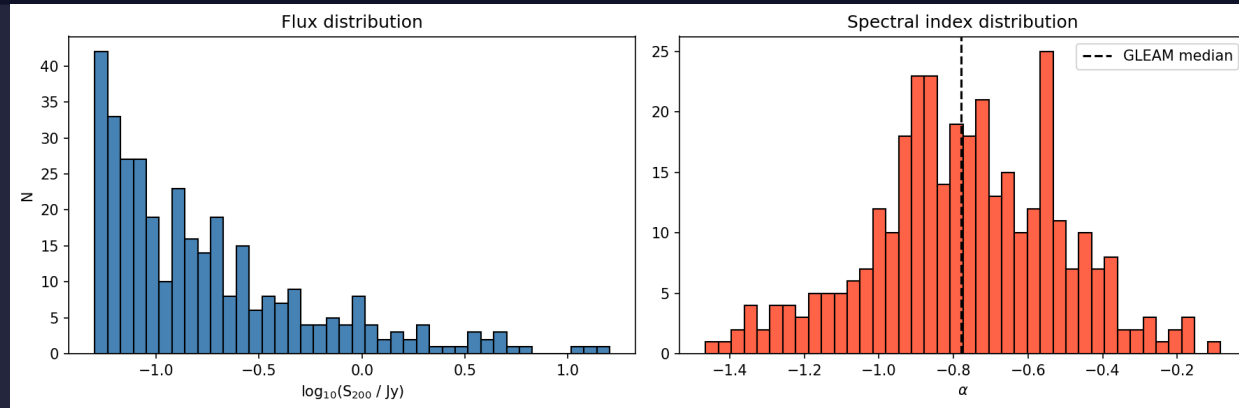


Figure 2: **Synthetic catalogue:** GLEAM-like flux and spectral-index distributions.

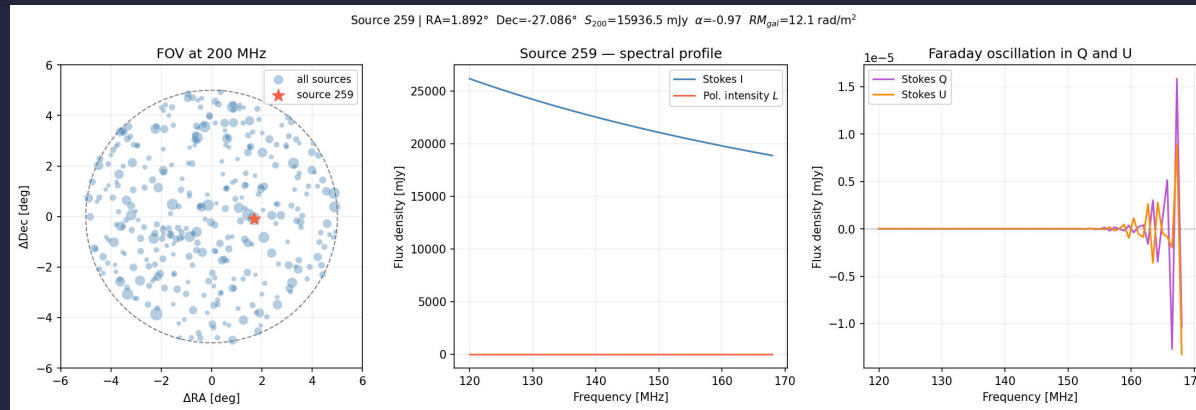


Figure 3: One realisation — 5° field at 200 MHz, per-source spectral profile, and the Faraday-rotated Q/U.

Polarization model

Simple model (intrinsic + galactic components)

$$P(\lambda^2) = p_0 \times I \times \exp(2i(\chi_0 + \phi_{\text{RM}}\lambda^2)) \times \mathcal{D}(\lambda^2)$$

where

- ϕ_0 : intrinsic fractional polarization
- χ_0 : intrinsic polarization angle
- $\phi_{\text{RM}} = \text{RM}_{\text{src}} + \text{RM}_{\text{gal}}$: rotation measure

The galactic term, we fetch from [Hutschenreuter+20](#) measurements

- $\mathcal{D}(\lambda^2)$: depolarization function

The Q and U terms are the real and imaginary parts of $P(\lambda^2)$. The V term is in almost all cases 0.

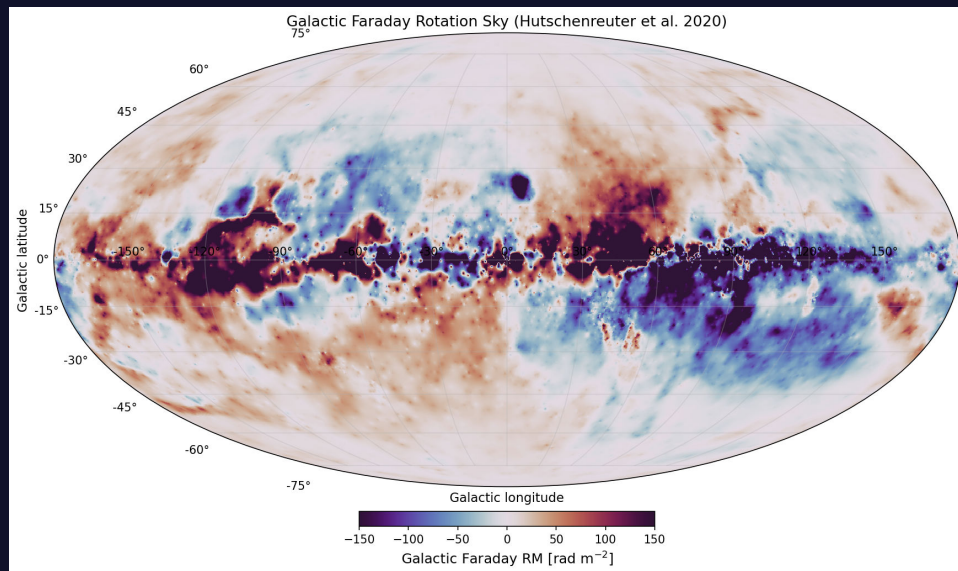


Figure 4: Galactic Faraday Rotation by [Hutschenreuter+20](#).

A metric gap

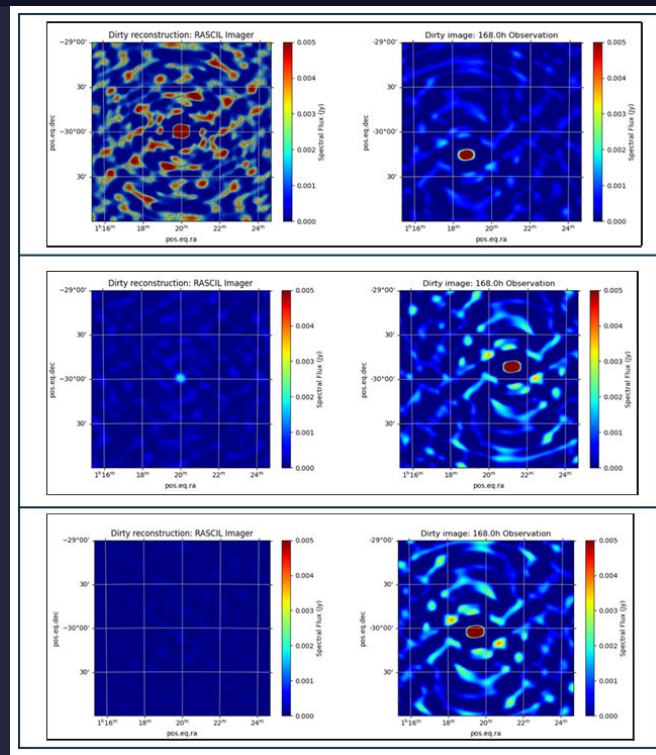
200 simulations, centred sources, 50 epochs

Model	PSNR (F)	PSNR (I)	MSSIM (I)
3D patch ViT	9.09	26.8	0.180
3D slice ViT	25.2	31.5	0.530
3D U-Net	27.4	32.4	0.683

1000 simulations, offset sources, 150 epochs

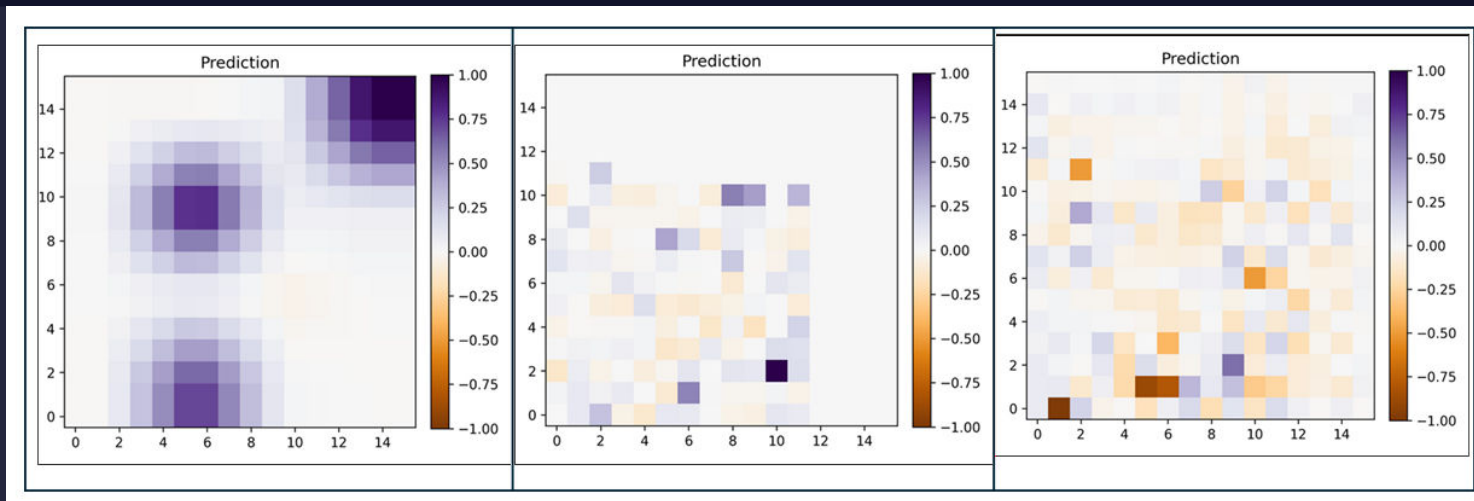
Model	PSNR (F)	PSNR (I)	MSSIM (I)
3D patch ViT	6.87	23.6	0.120
3D slice ViT	29.5	35.7	0.680
3D U-Net	29.1	33.7	0.640

The U-Net won every metric while reconstructing almost nothing.



Reconstruction (left) vs. truth (right): patch ViT, slice ViT, U-Net.

Inductive bias must match the physics



Predicted centre slices in the $u-v$ plane. Left: patch ViT. Centre: slice ViT. Right: U-Net.

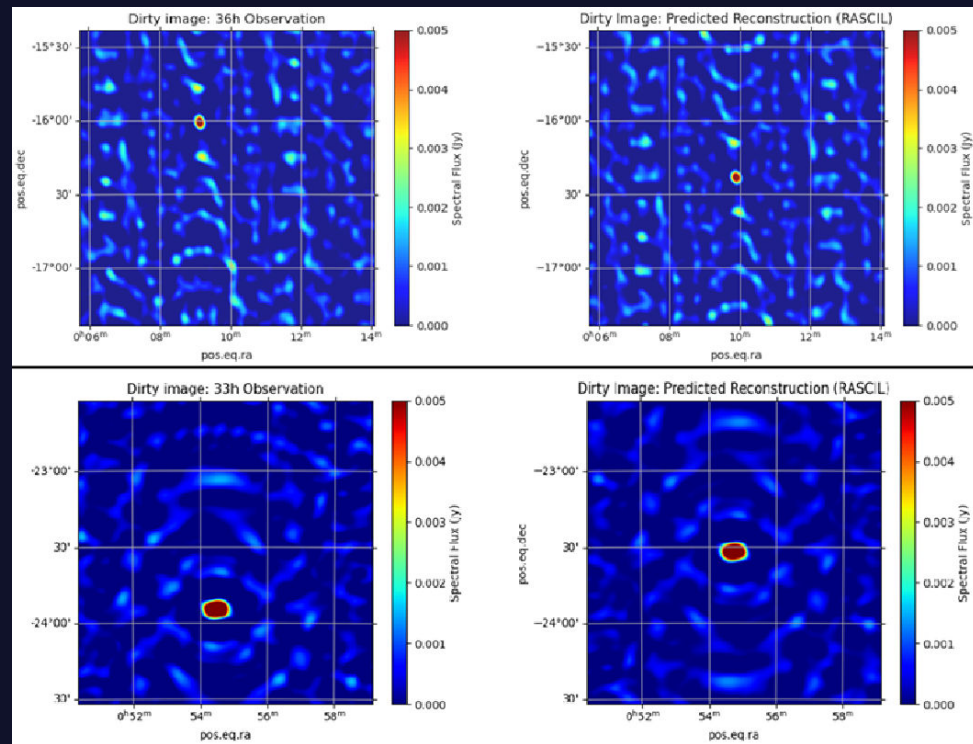
- **Patch tokenisation** assumes neighbouring pixels form a coherent spatial neighbourhood. In the Fourier domain adjacent frequels need not be adjacent baselines — so it hallucinated smooth, image-like blobs
- **Convolutions** assume locality and smoothness — also wrong for globally-coupled visibility data
- **Slice tokenisation** — one $u-v$ slice per token, attention across time — matches the real structure, and wins once trivial solutions stop being rewarded

Representations are important: learning latents

Latent compression (200 epochs)

Model / condition	PSNR (I)	MSSIM (I)
AE (f16)	20.3	0.00
AE (f16) + amplitude/phase	24.4	0.08
AE (f8)	34.7	0.77
AE (f8) + amplitude/phase	41.9	0.92

- The latent respects the spatial-frequency coordinates better, and produces **realistic** emission



Latent *forecasting* still has problems — the learning objective splits the visibility into amplitude and phase, where amplitudes are prioritized without intervention. Phase contains positional information.

Lessons learned

"Physics-informed" design

- **Spatiotemporal structure is non-negotiable**

Models need spatial coherence AND temporal continuity (3D); 2D formulations fail

- **Slice-based ViT respects uv-plane topology**

Attention over full slices

- **Amplitude-weighting prevents collapse**

generic MSE doesn't work, scaled residuals yield better results

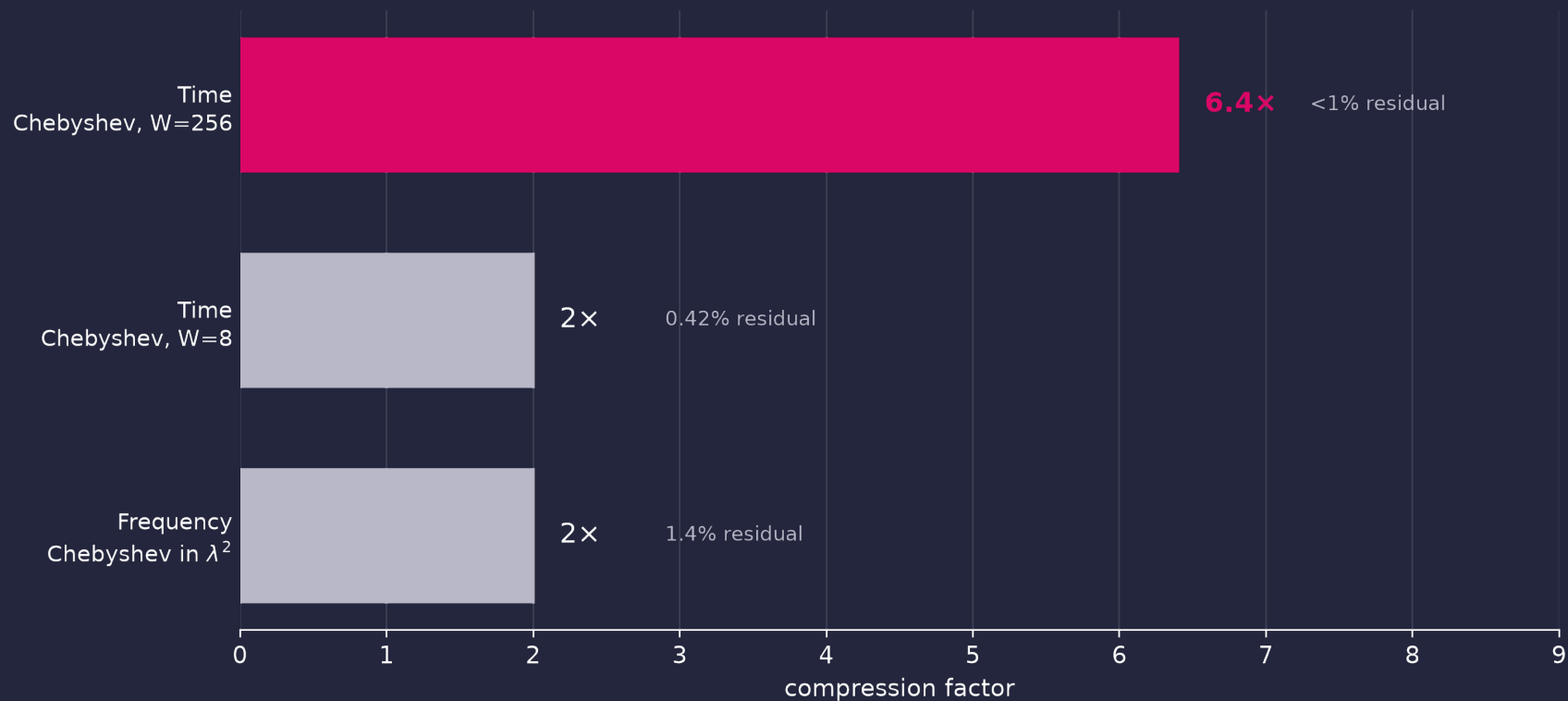
Challenge

- **Scalability bottleneck**

Compression is necessary to make AA* and AA4 feasible

Compression — measured, axis by axis

- **Time** — Chebyshev, our best axis
- **Frequency** — compression $\times 2$ but larger residual



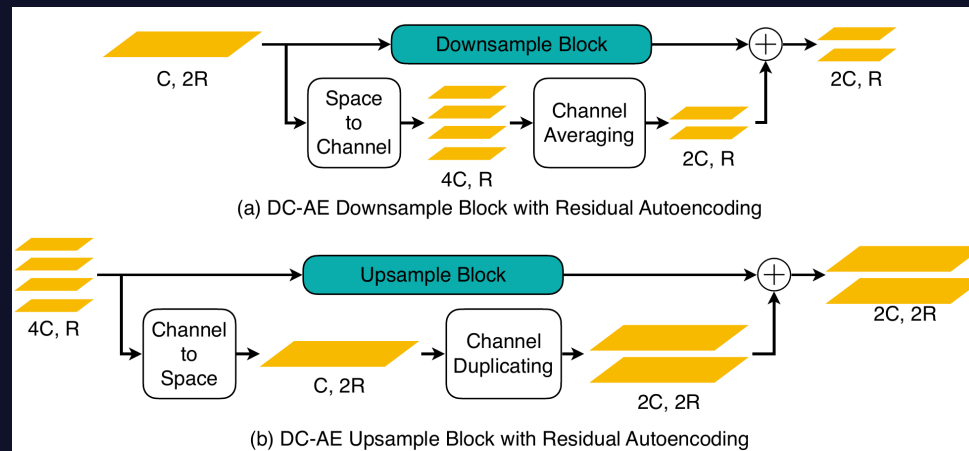
Measured on 3600 timesteps $\times$ 64 channels $\times$ 2278 baselines $\times$ 4 pol. Time $\times$ frequency combined: $\approx 4\times$ at W=8, $\approx 13\times$ at W=256.

Deep compression

Deep Compression Autoencoder

(MIT Han Lab, [arXiv:2410.10733](https://arxiv.org/abs/2410.10733)) — the encoder behind NVIDIA's **SANA**:

- spatial compression **32×**, **64×**, **128×**, where an SD-VAE stalls at **8×**
- `dc-ae-f32c32` : $512 \times 512 \rightarrow 16 \times 16 \times 32$
- **Residual autoencoding** — learn residuals on space-to-channel transformed features
- **Decoupled high-resolution adaptation** — a three-phase training schedule
- **19.1×** inference and **17.9×** training speed-up on ImageNet-512, at better FID than SD-VAE-f8



Residual autoencoding (Chen+25, Fig. 4). A *non-parametric* space-to-channel shortcut is added around every down- and upsample block, so the learned module only has to fit the **residual**:
 $H \times W \times C \rightarrow \frac{H}{2} \times \frac{W}{2} \times 4C$, channel-averaged back to $2C$. Nothing is discarded — resolution is *traded for channels*.

That shortcut is what makes 32–128× compression *trainable* at all.

1. **The representation is important** — amplitude/phase: an AE at f8 reaches **0.92** MSSIM (I).
2. **Inductive bias must match the sampling** — patch and convolutional priors assume an image; the (u, v) plane is not one.
3. forecasting favours **amplitude**, but *position lives in phase*.

Next: phase-aware objectives · deep compression autoencoders · latent forecasting on a video-prediction backbone

Slides (online version)



on phdenzel.github.io

Contact

Philipp Denzel

Centre for Artificial Intelligence, ZHAW

Technikumstrasse 71, Winterthur 8400

denp@zhaw.ch

