HEAL-SWIN:
A Vision Transformer On The Sphere

Oscar Carlsson[1], Jan E. Gerken[1], Hampus Linander,
Heiner Spieß, Fredrik Ohlsson, Christoffer Petersson,
Daniel Persson

Division of algebra and geometry, Department of Mathematical Sciences
Chalmers University of Technology and Gothenburg University

Accepted as poster at CVPR 2024

[1]Equal contribution

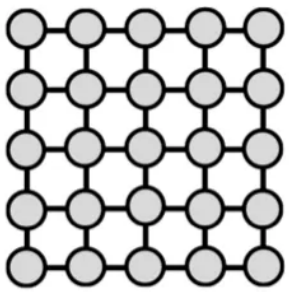
Outline

- Geometry
- Transformers
- The HEAL-SWIN model
- Experiments and results

Geometry

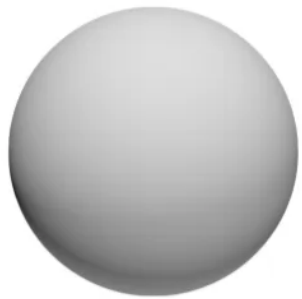
Geometric deep learning

Grids



Euclidean samples,
e.g. image

Groups



Homogenous spaces
with global symmetries,
e.g. sphere

Graphs

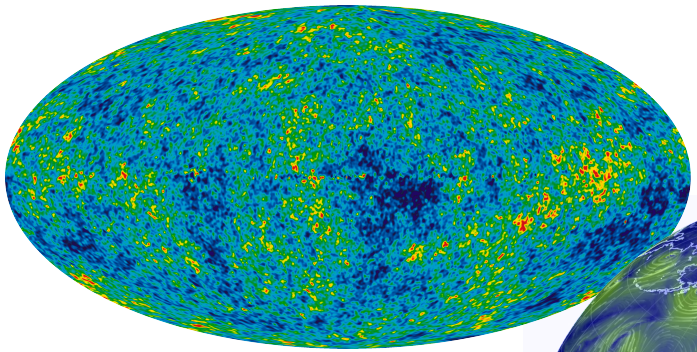
Nodes and connections,
e.g. social network

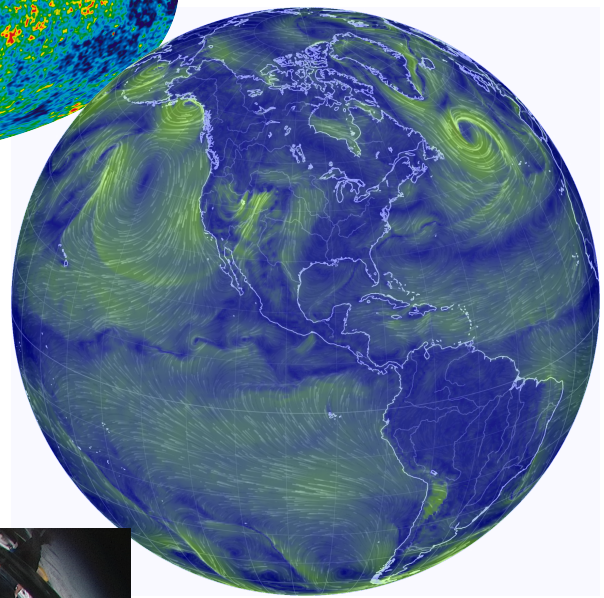
Geodesics & Gauges

Manifolds,
e.g. 3D mesh

From Bronstein et al. (2021)

Spherical geometry

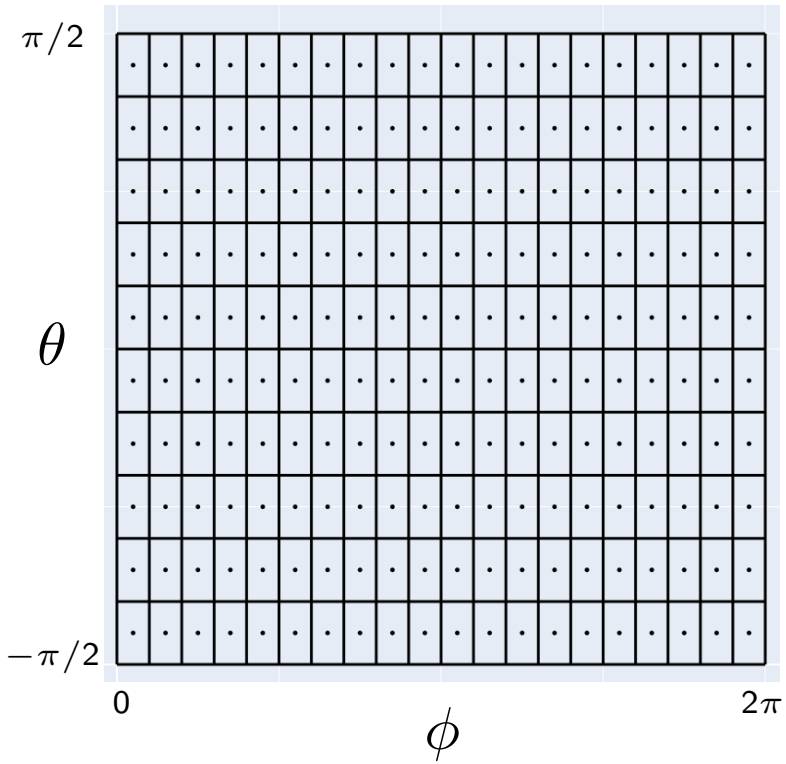




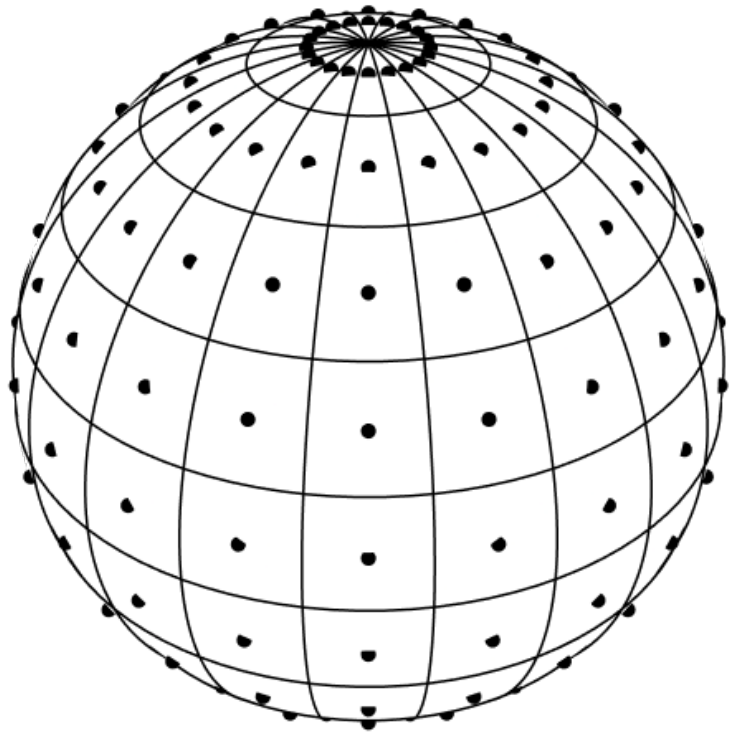
- Astrophysics
- Weather modelling
- Chemistry
- Autonomous vehicles

Driscoll Healy (Equiangular) representation

View in θ, ϕ space



3D view



Different approaches to spherical data

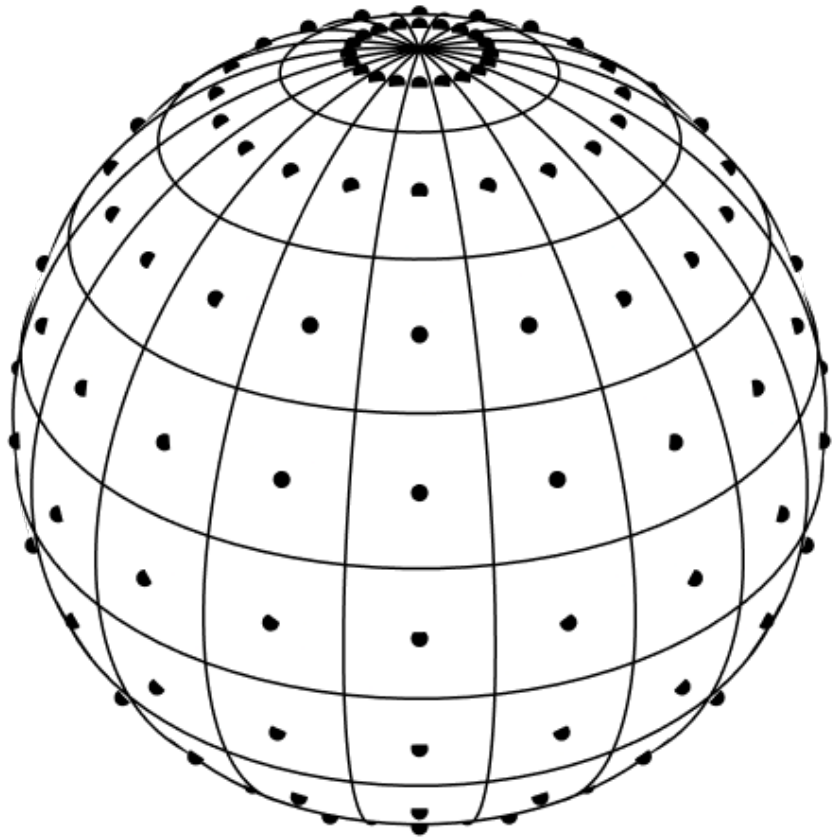
- Local flat approximations of the sphere

Different approaches to spherical data

- Local flat approximations of the sphere
- Fourier based CNNs

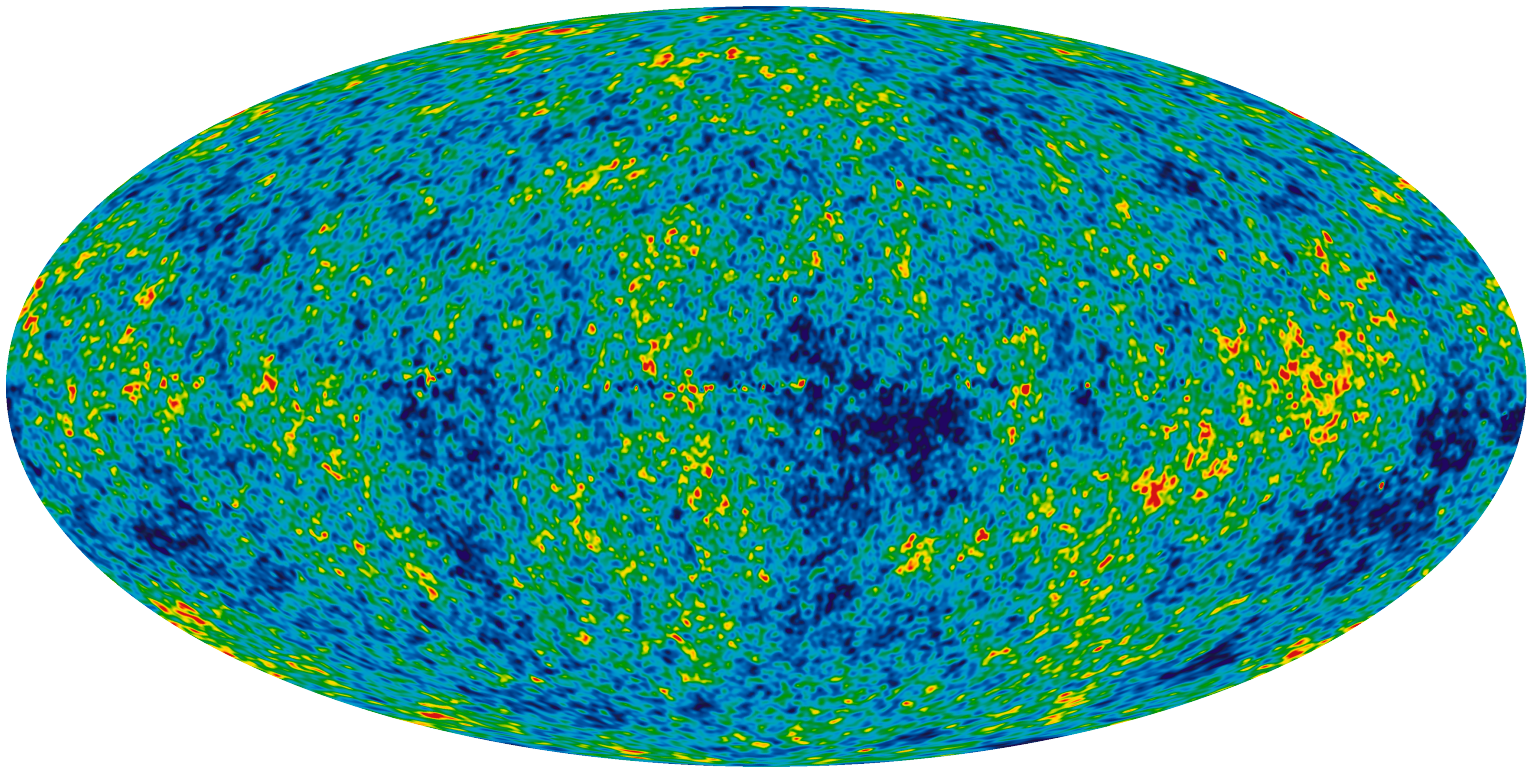
Different approaches to spherical data

- Local flat approximations of the sphere
- Fourier based CNNs

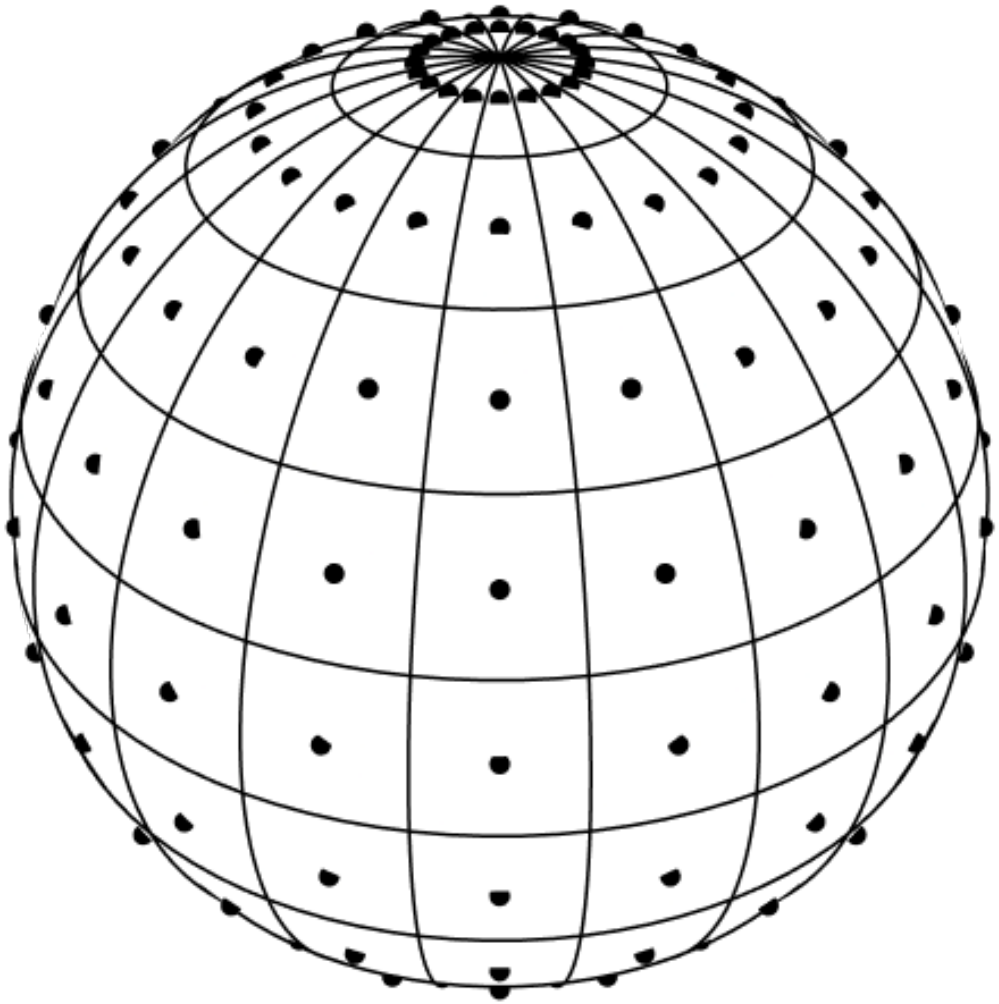


HEALPix

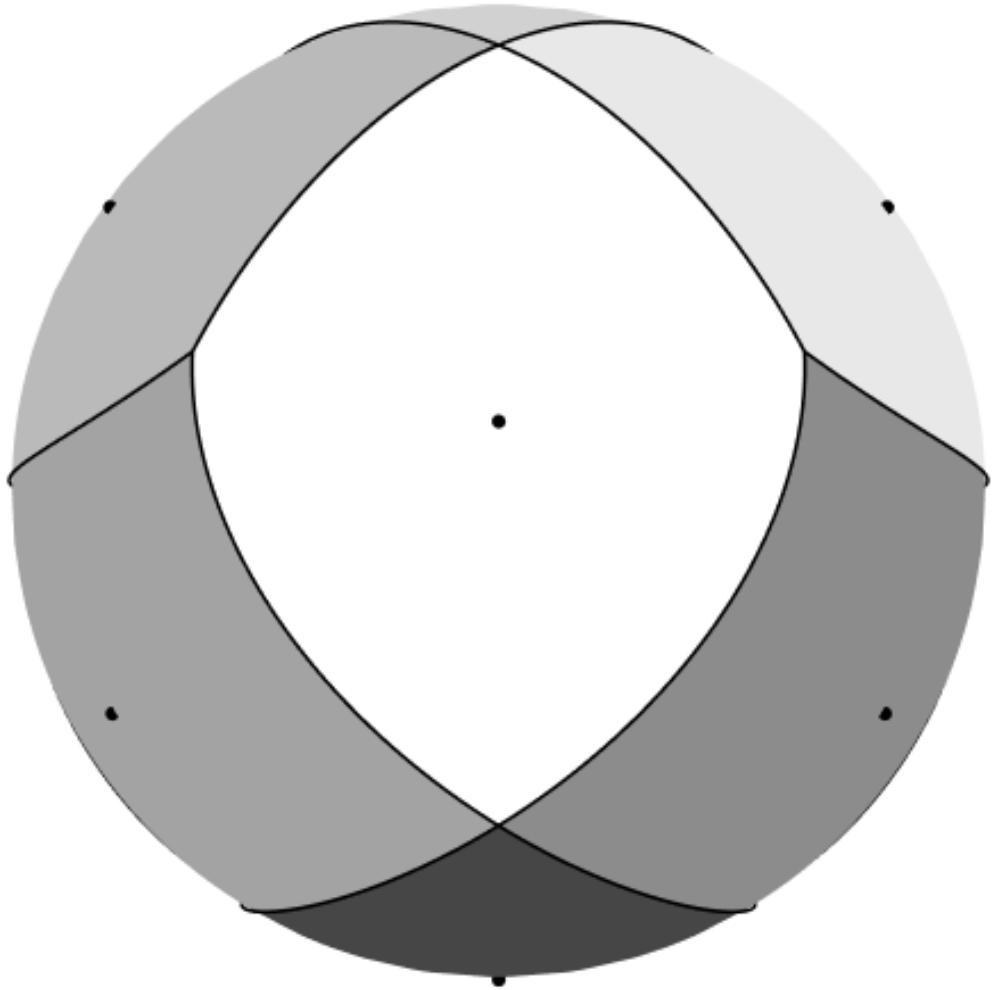
HEALPix background (Gorski et al., 1998)



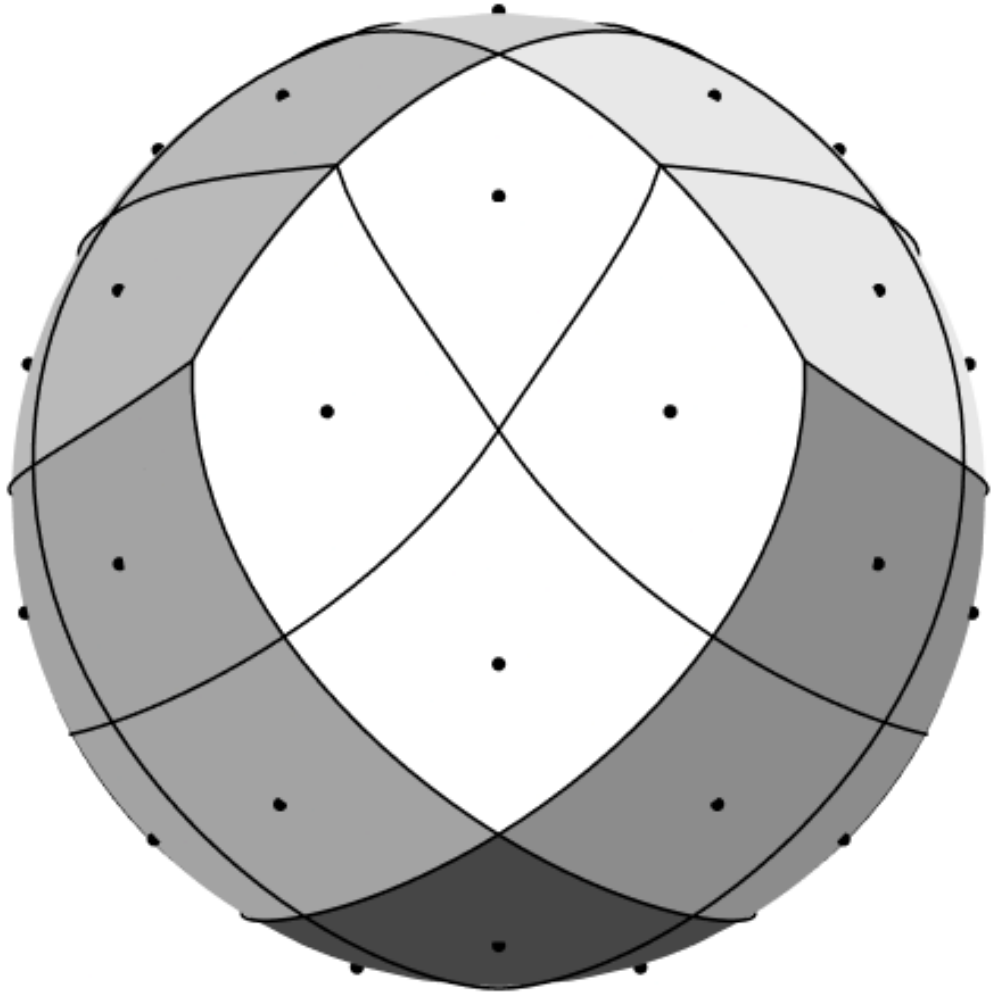
HEALPix construction, instead of



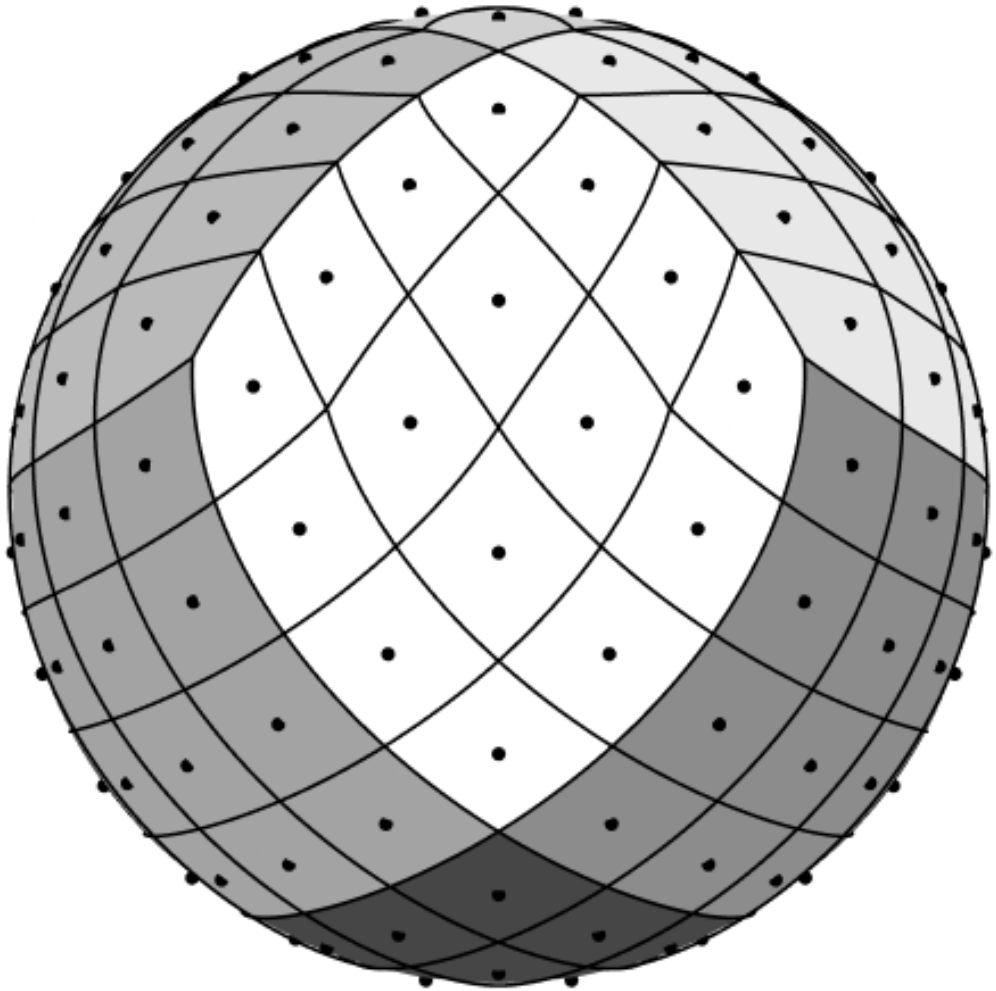
HEALPix construction, do



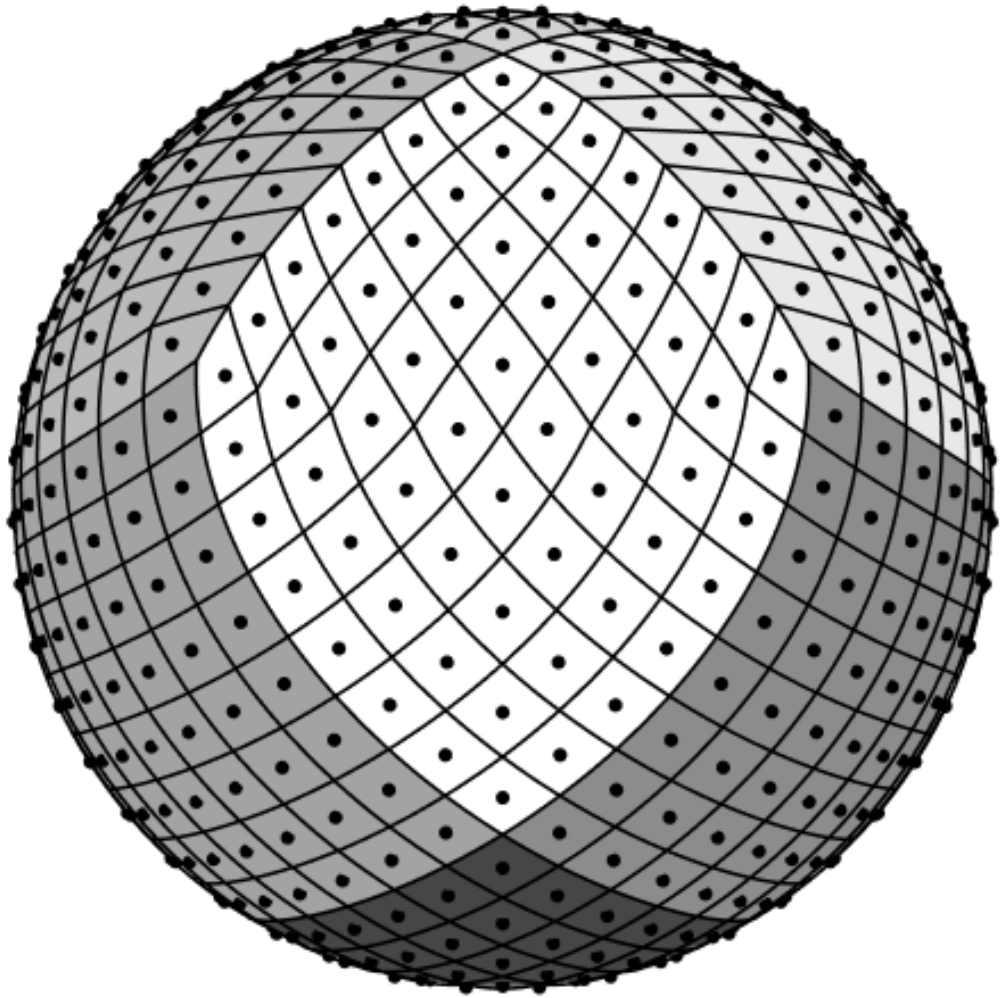
HEALPix construction, do



HEALPix construction, do

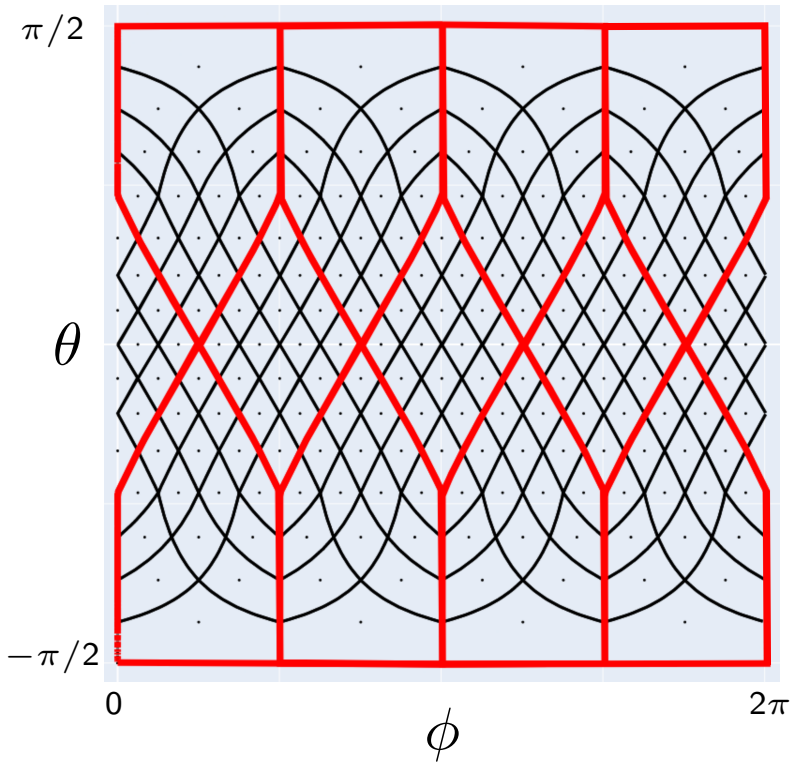


HEALPix construction, do

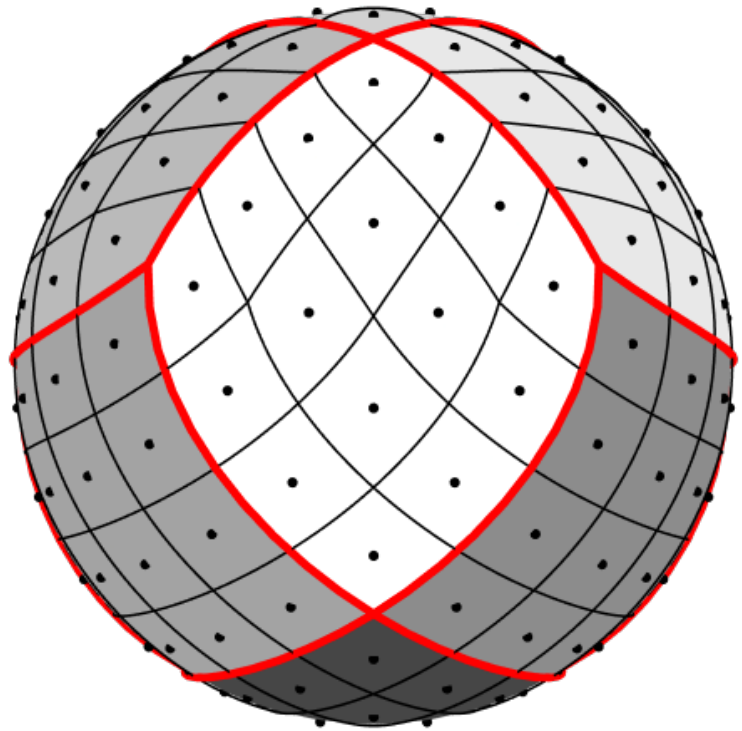


HEALPix visualisation for nside= 4

View in θ, ϕ space



3D view

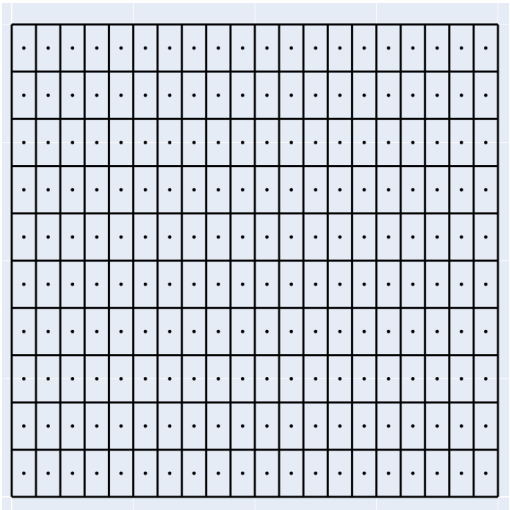


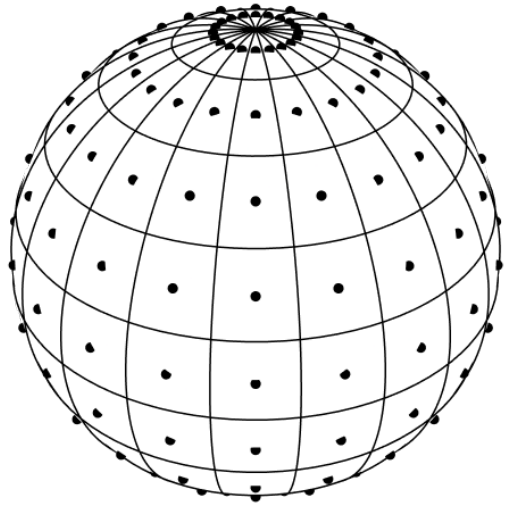
HEALPix vs Driscoll Healy

View in θ, ϕ space

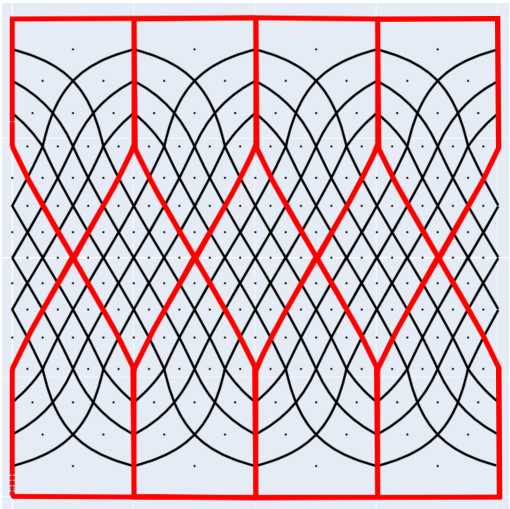
3D view

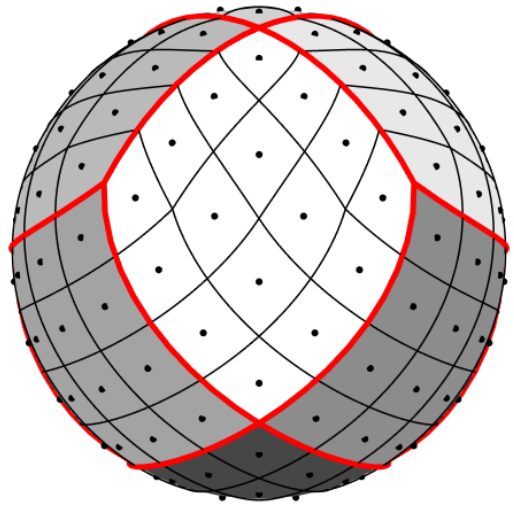
Driscoll Healy





HEALPix





Transformers

Transformers and attention weights

- Given a set of inputs (tokens) $\{x_i \in \mathbb{R}^{d_{in}}\}_{i=1,\ldots,N}$ (think, e.g., the set of pixels in an image all having values in $\mathbb{R}^3$ for an RGB image, or words in a sentence)
- Embed these into some space $x'_i \in \mathbb{R}^d$
- (Self) Attention weight between x'_i and x'_j is

$$\alpha(i,j) = \frac{\langle Q_i, K_j \rangle}{\sqrt{d}}$$

where $Q_i = W^Q x'_i$, $K_j = W^K x'_j$, hence $\alpha \in \mathbb{R}^{N \times N}$.

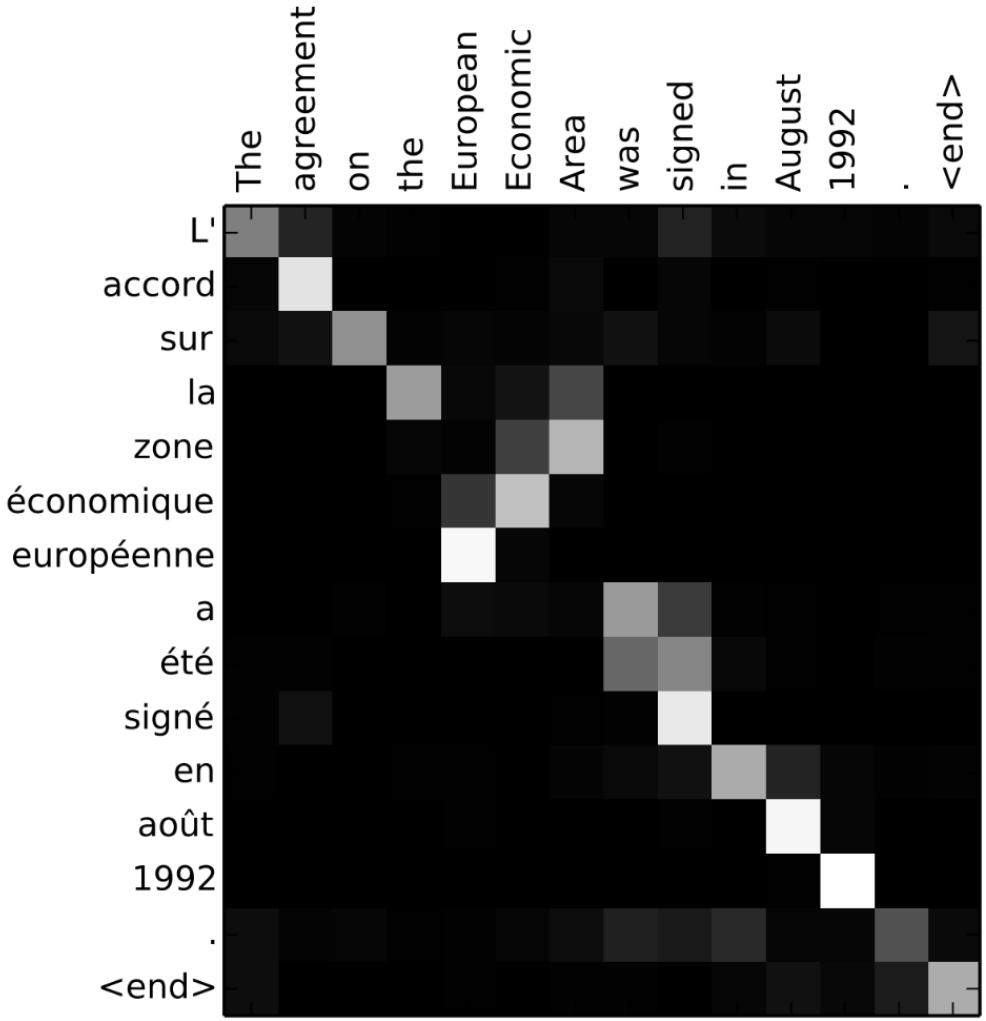
- Output

$$z_i = \sum_j \text{softmax}(\alpha(i,:))_j V_j$$

where $V_j = W^V x'_j$

- Scales quadratically with respect to number of tokens!

Attention visualisation



Attention-matrix heatmap

Bahdanau, et al. 2015. Neural machine translation by jointly learning to align and translate. In Proc. ICLR.

Vision transformers

- "Normal images" are quite large, $\sim 30 \cdot 10^6$ pixels
- Applying a transformer straight on the input is (somewhat) unfeasable
- Embed patches instead of pixels!

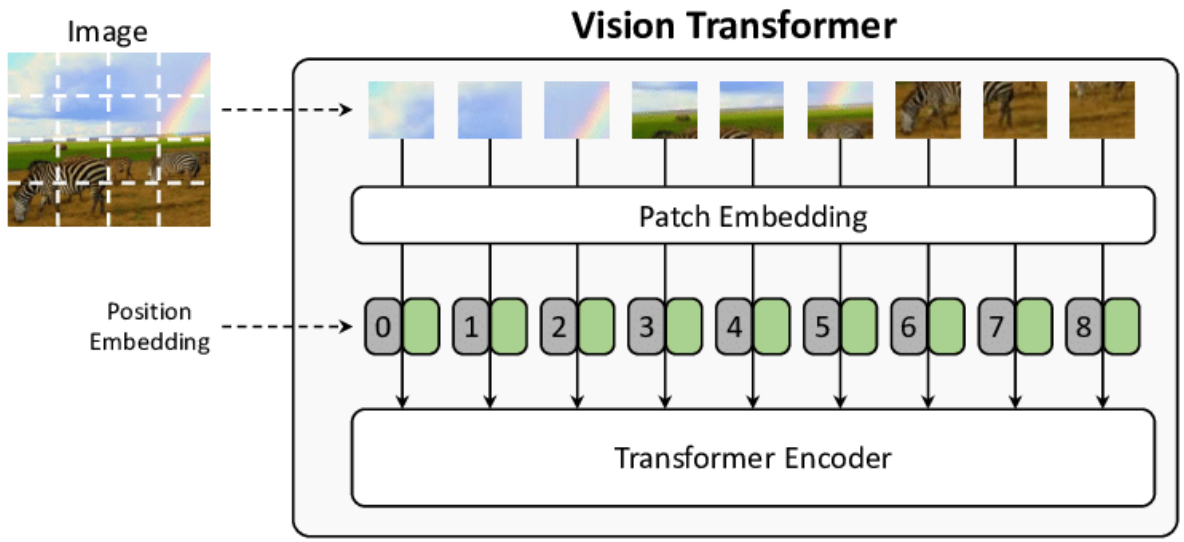
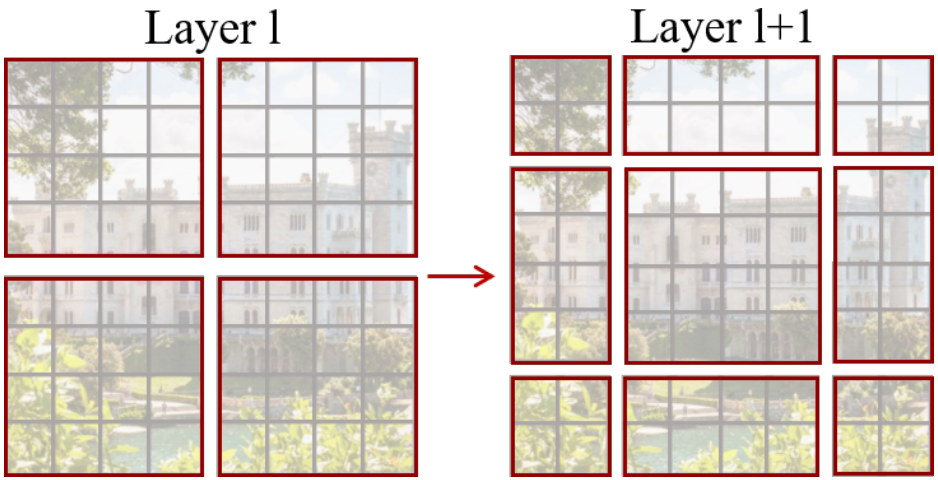


Figure from Stefanini et al. (2021)

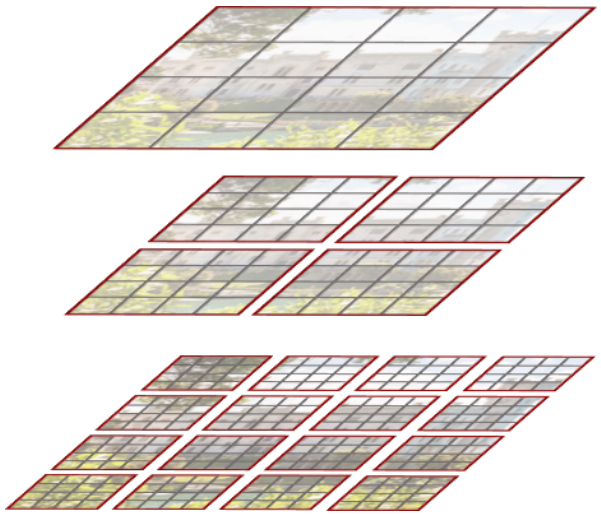
- Either attention over all patches or subgroups of pathes (windows)
- Lacks inductive biases!

The Shifting WINdow transformer (SWIN-transformer)

Shifting attention windows

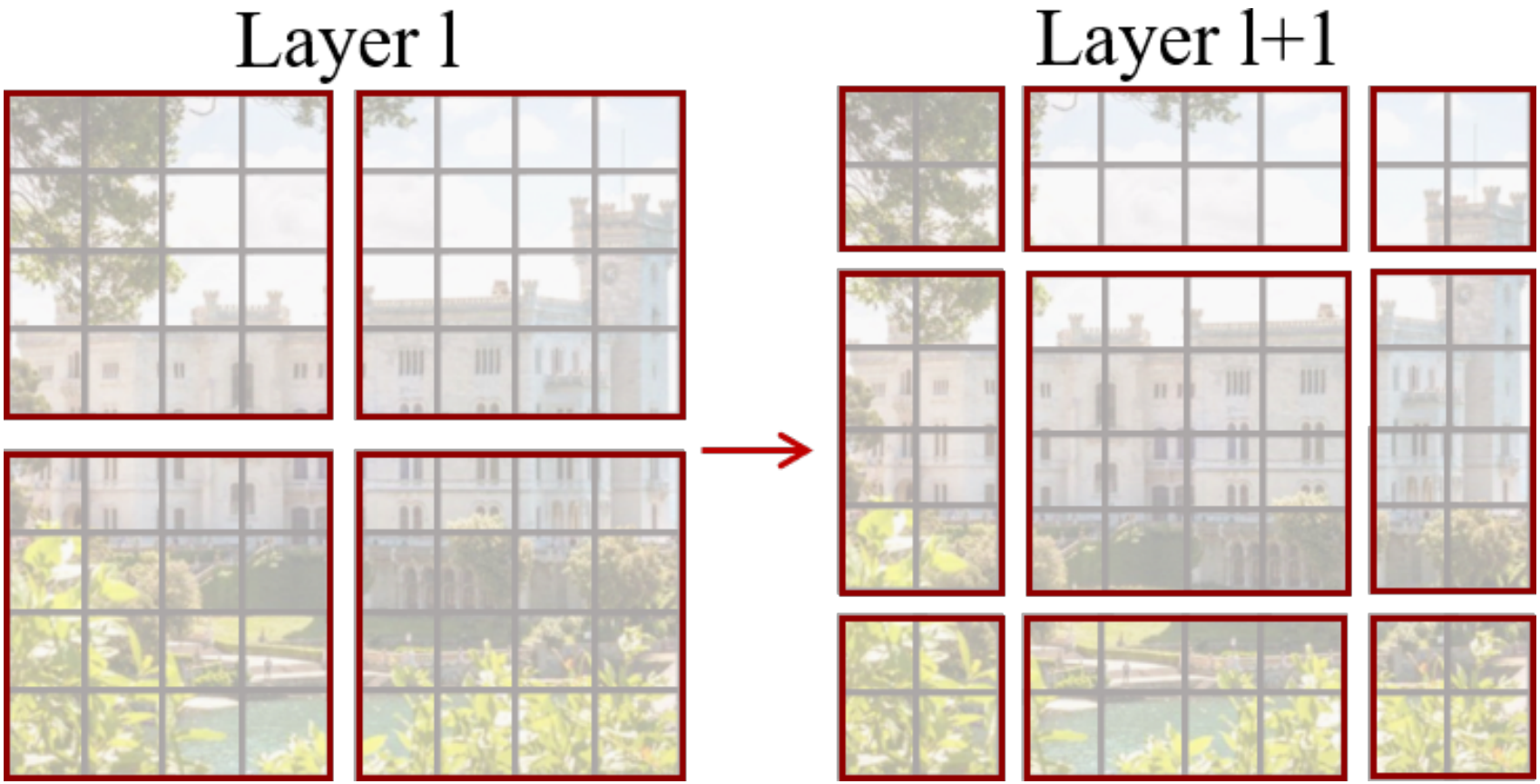


Patch merging

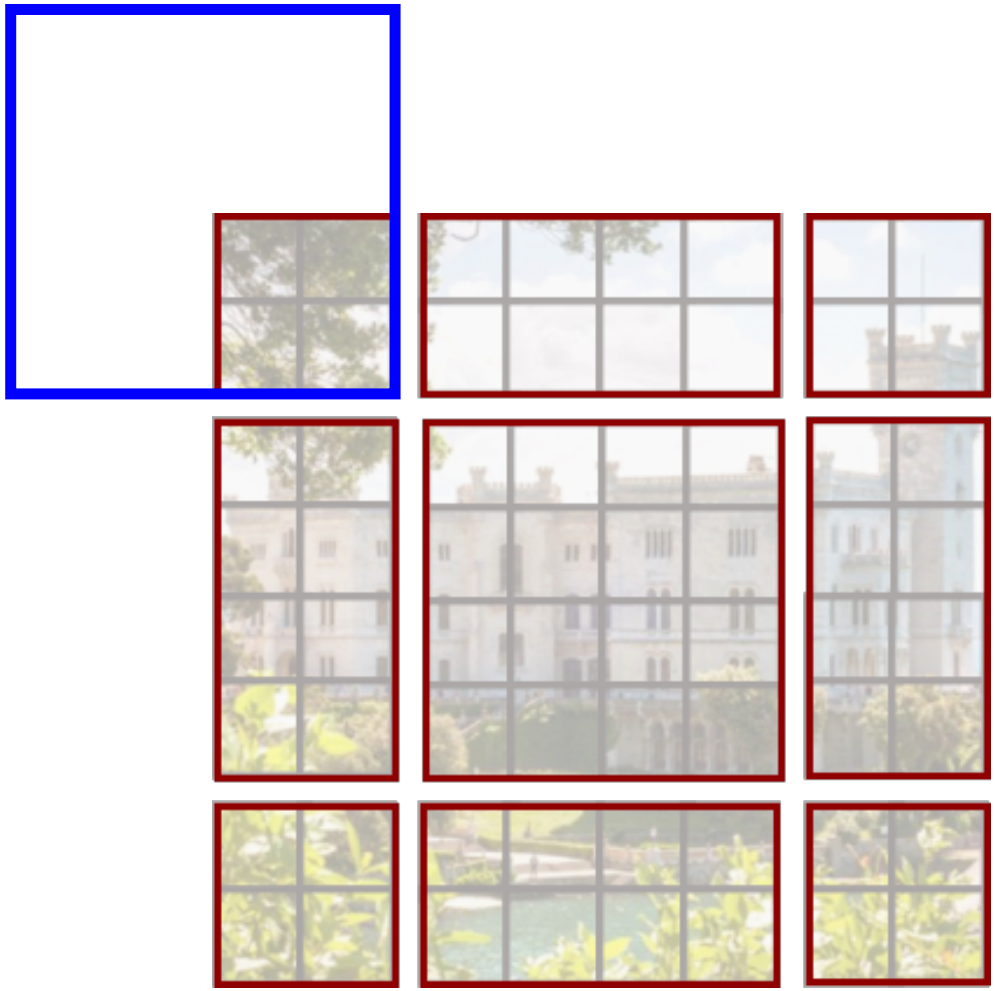


The SWIN transformer (Liu et al., 2021).

The SWIN window shift

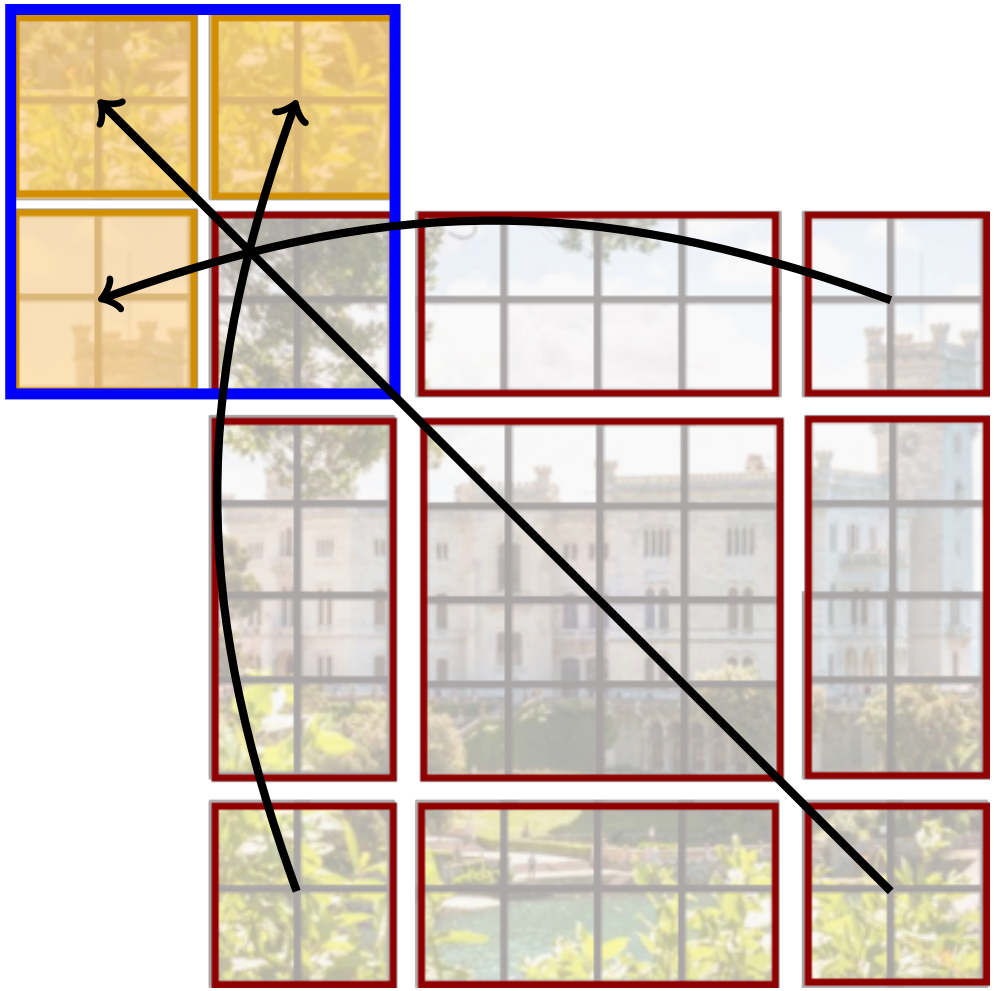


Partially filled attention windows

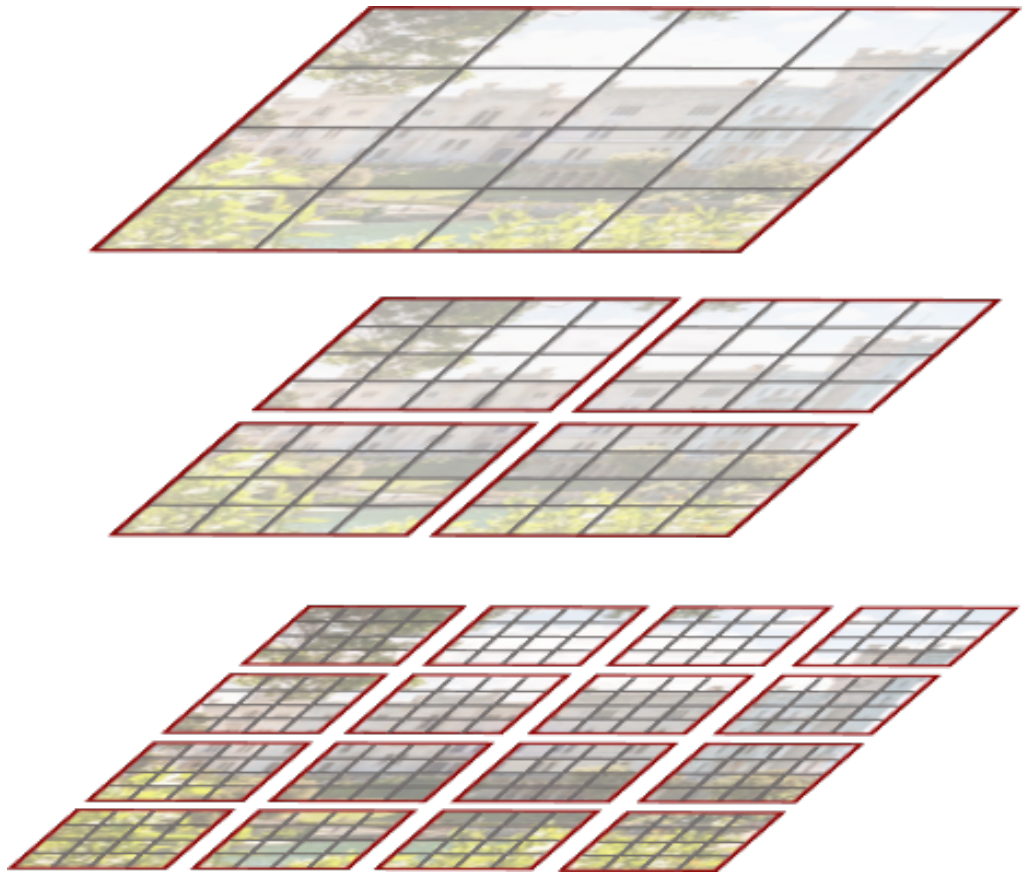


Partially filled attention windows

Partially filled attention windows



The SWIN patch merging

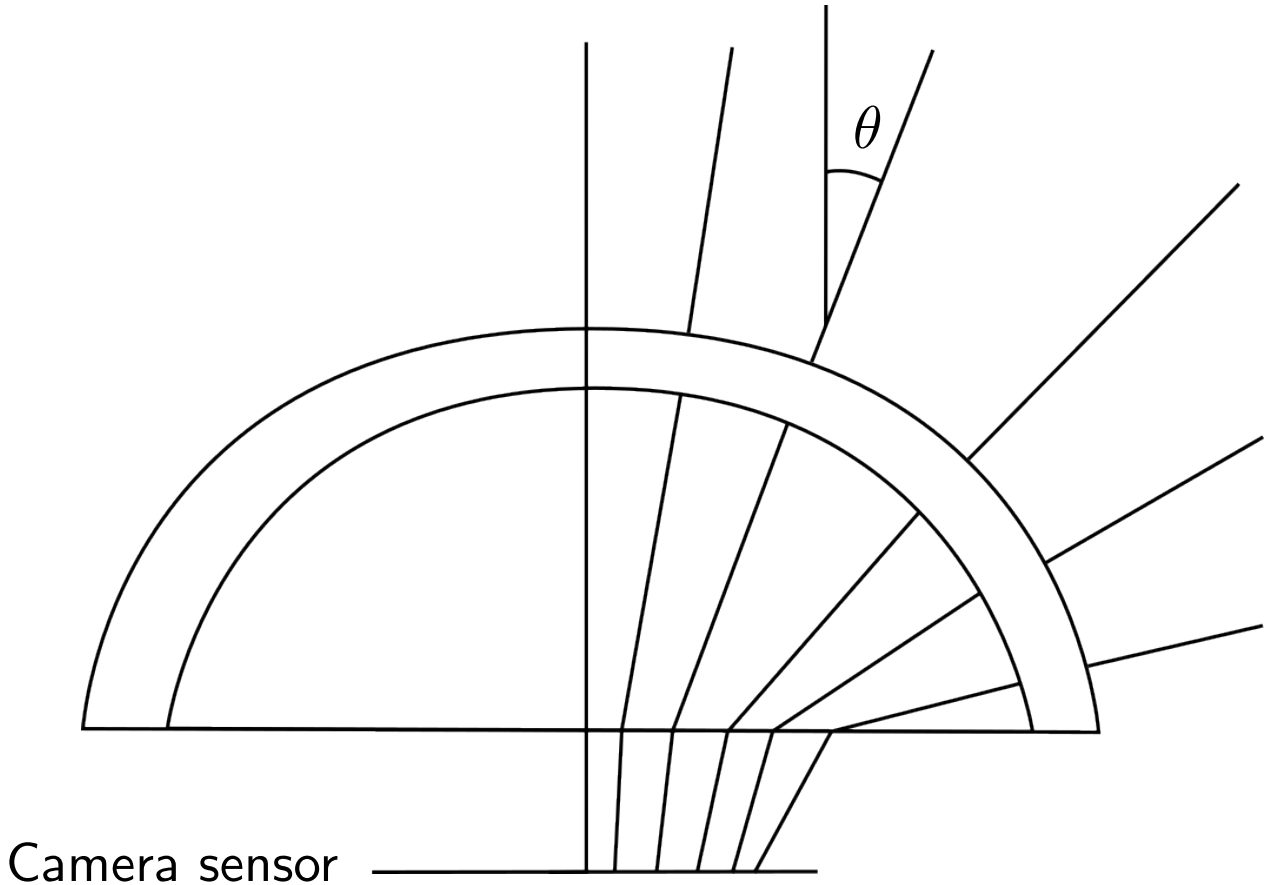


The HEAL-SWIN model

Central problem

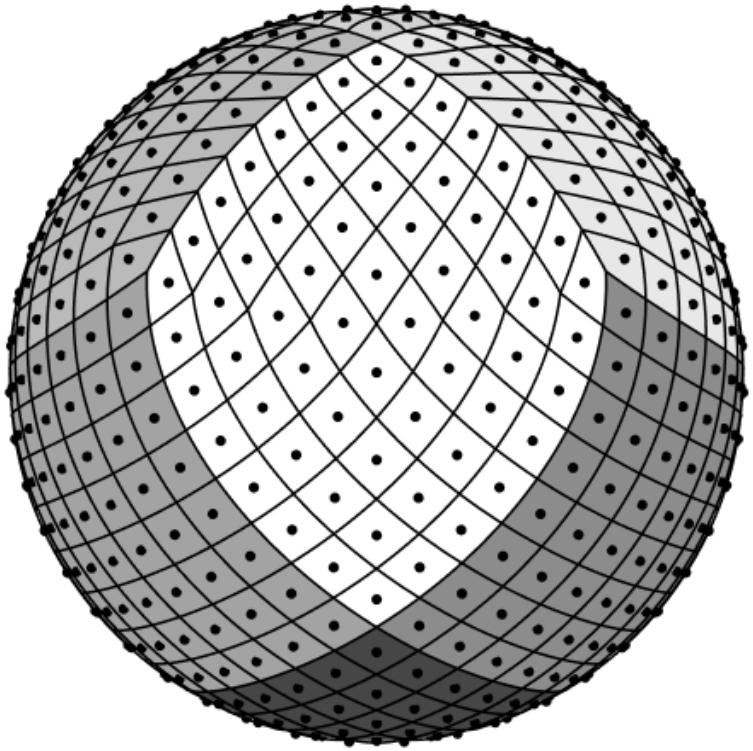
Example image from the WoodScape dataset
(Ramachandran et al., 2021)

Why spherical?

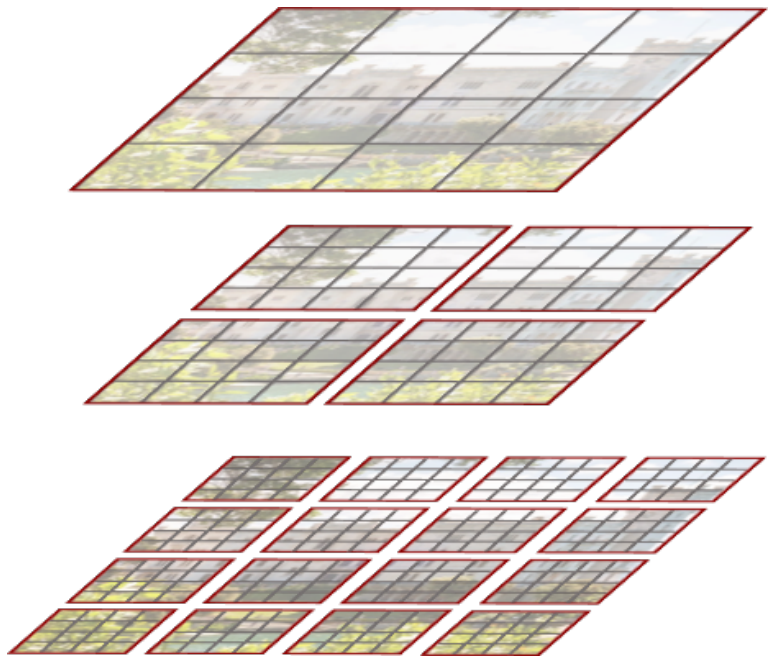


Schematic of a fisheye camera

HEAL-SWIN: HEALPix and the SWIN transformer



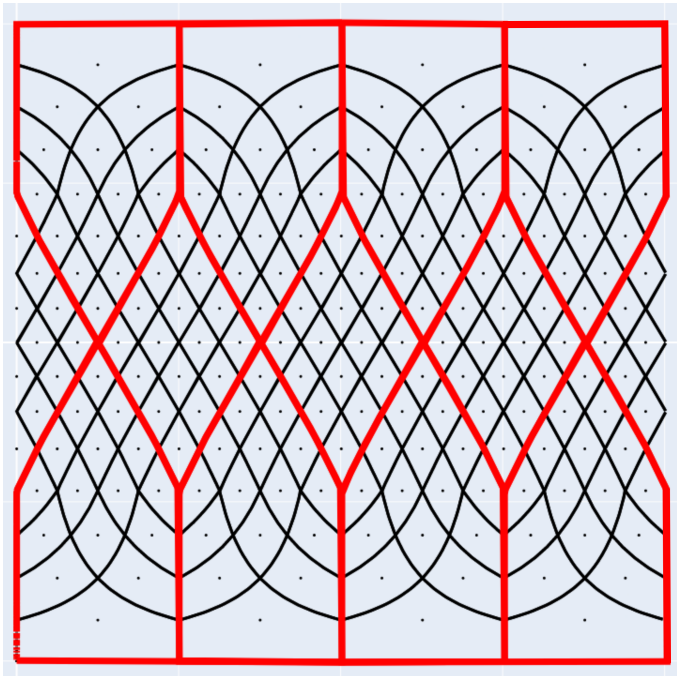
The HEALPix grid
(Gorski et al., 1998)



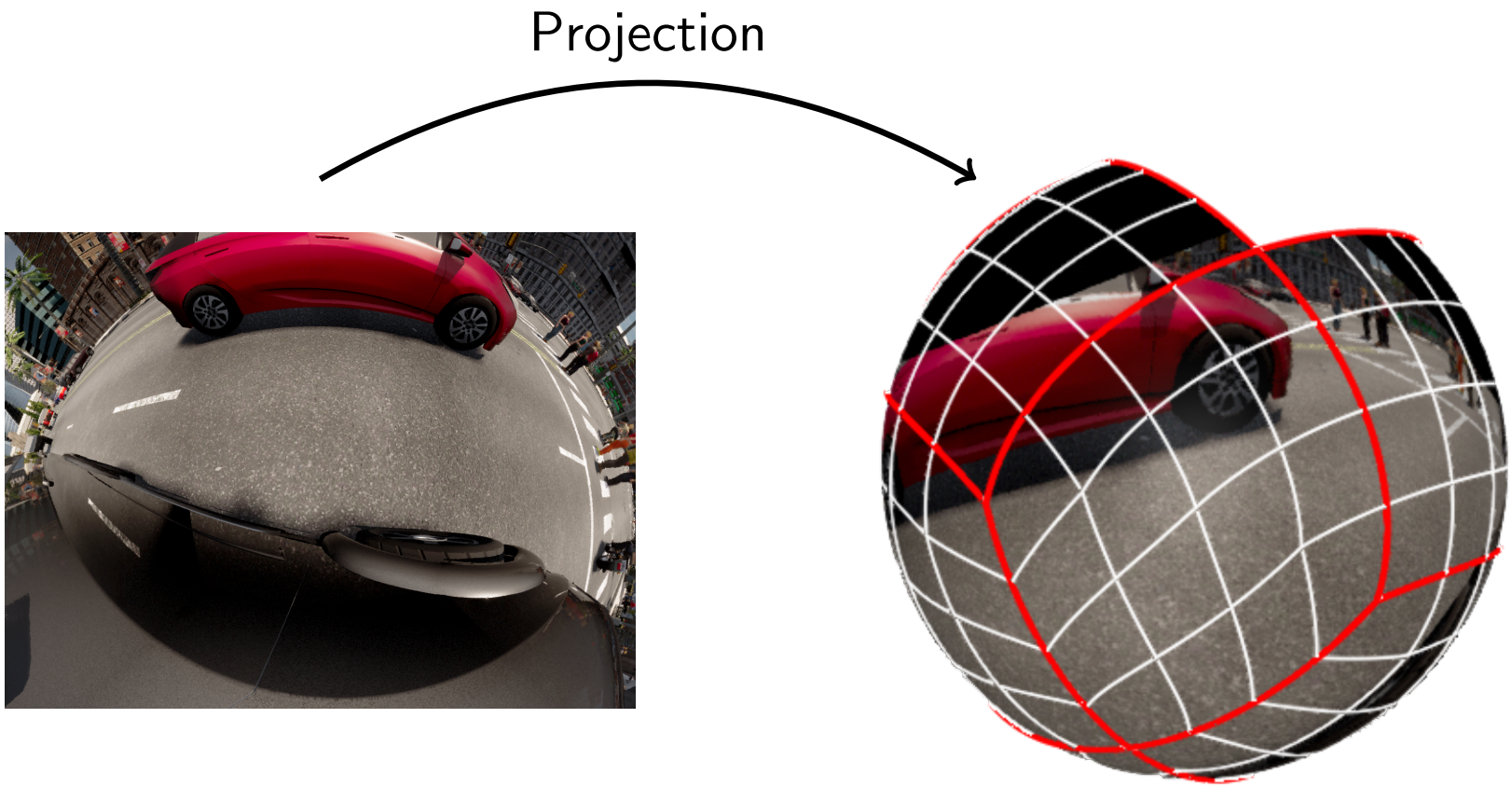
The SWIN transformer
(Liu et al., 2021)

Projecting images to HEALPix

Projecting images to HEALPix

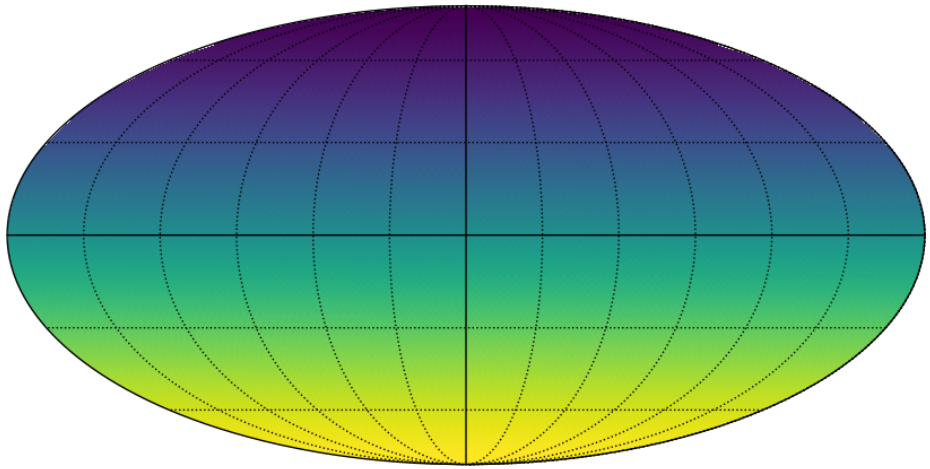


Projecting images to HEALPix



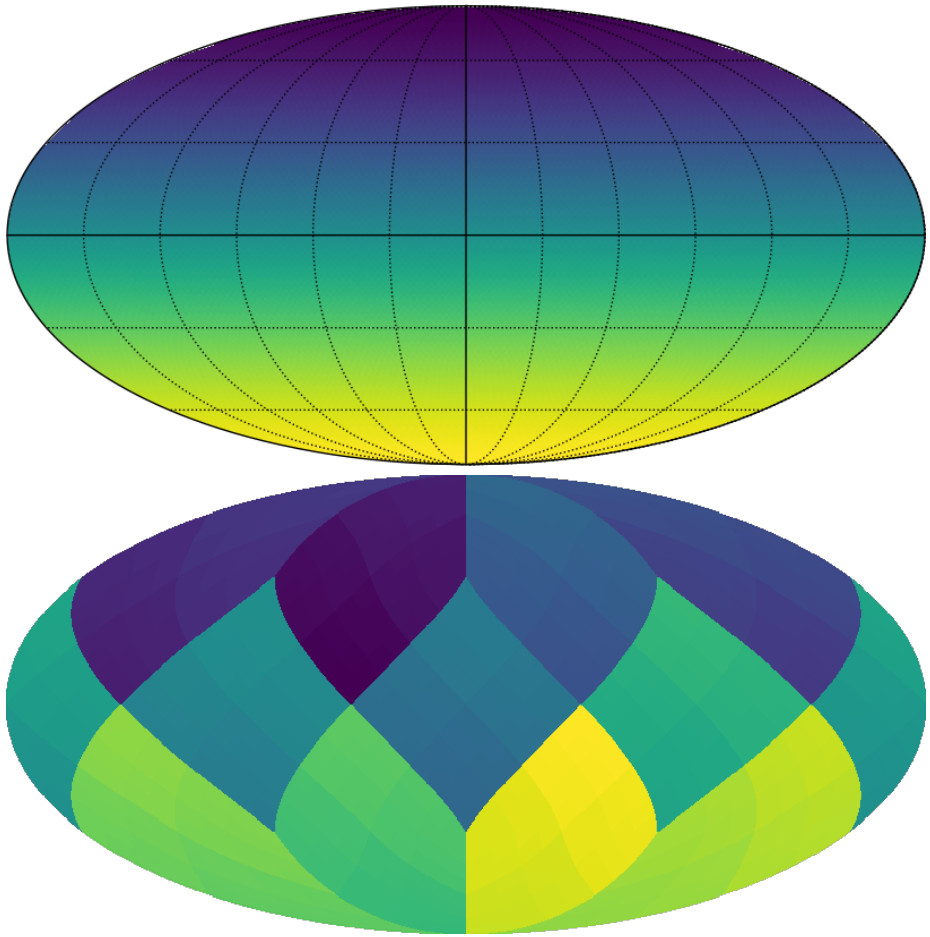
HEALPix array representations

- Ring

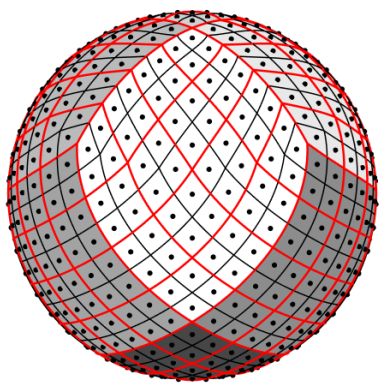


HEALPix array representations

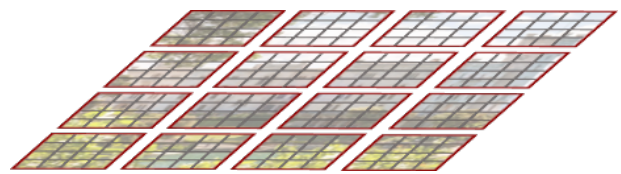
- Ring
- Nested



HEAL-SWIN patch merging

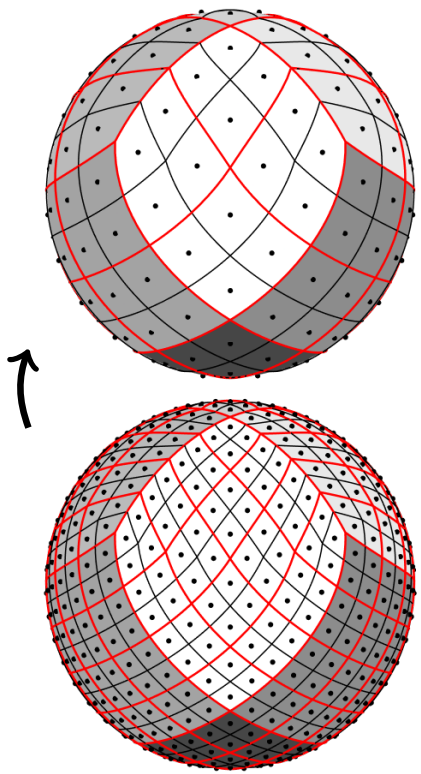


HEAL-SWIN

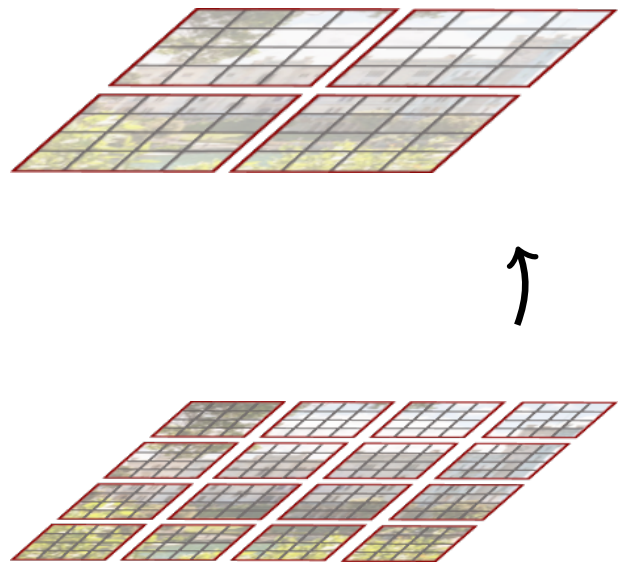


SWIN

HEAL-SWIN patch merging

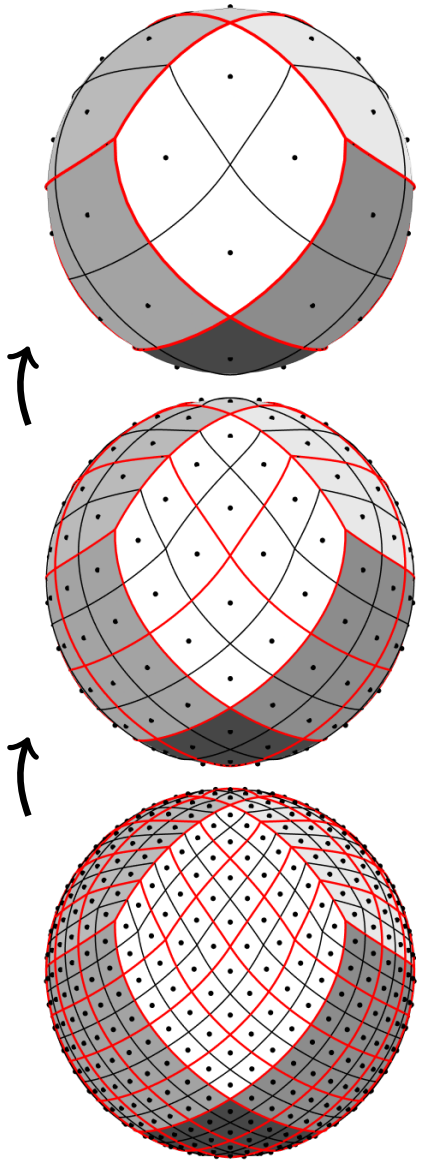


HEAL-SWIN

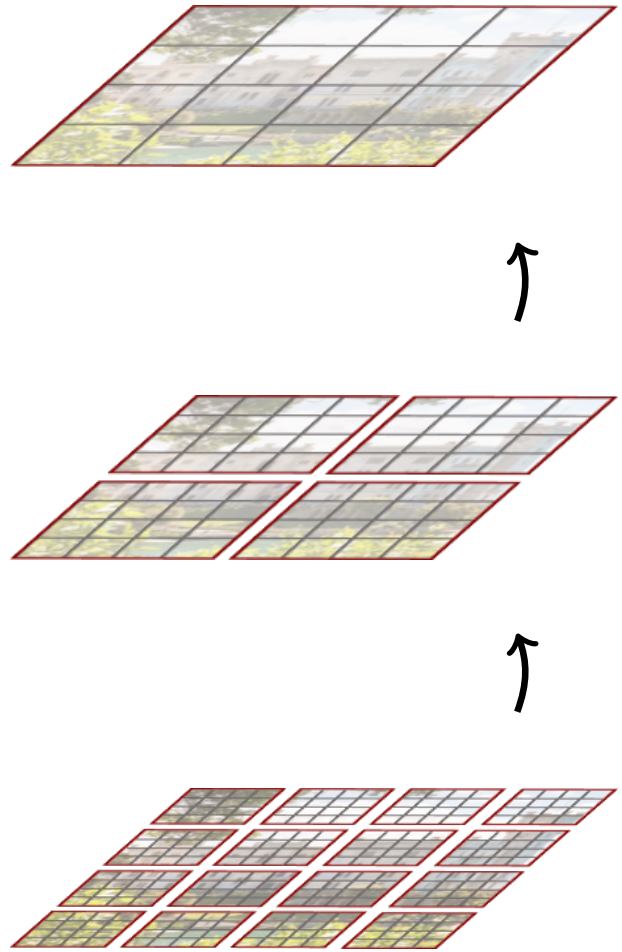


SWIN

HEAL-SWIN patch merging

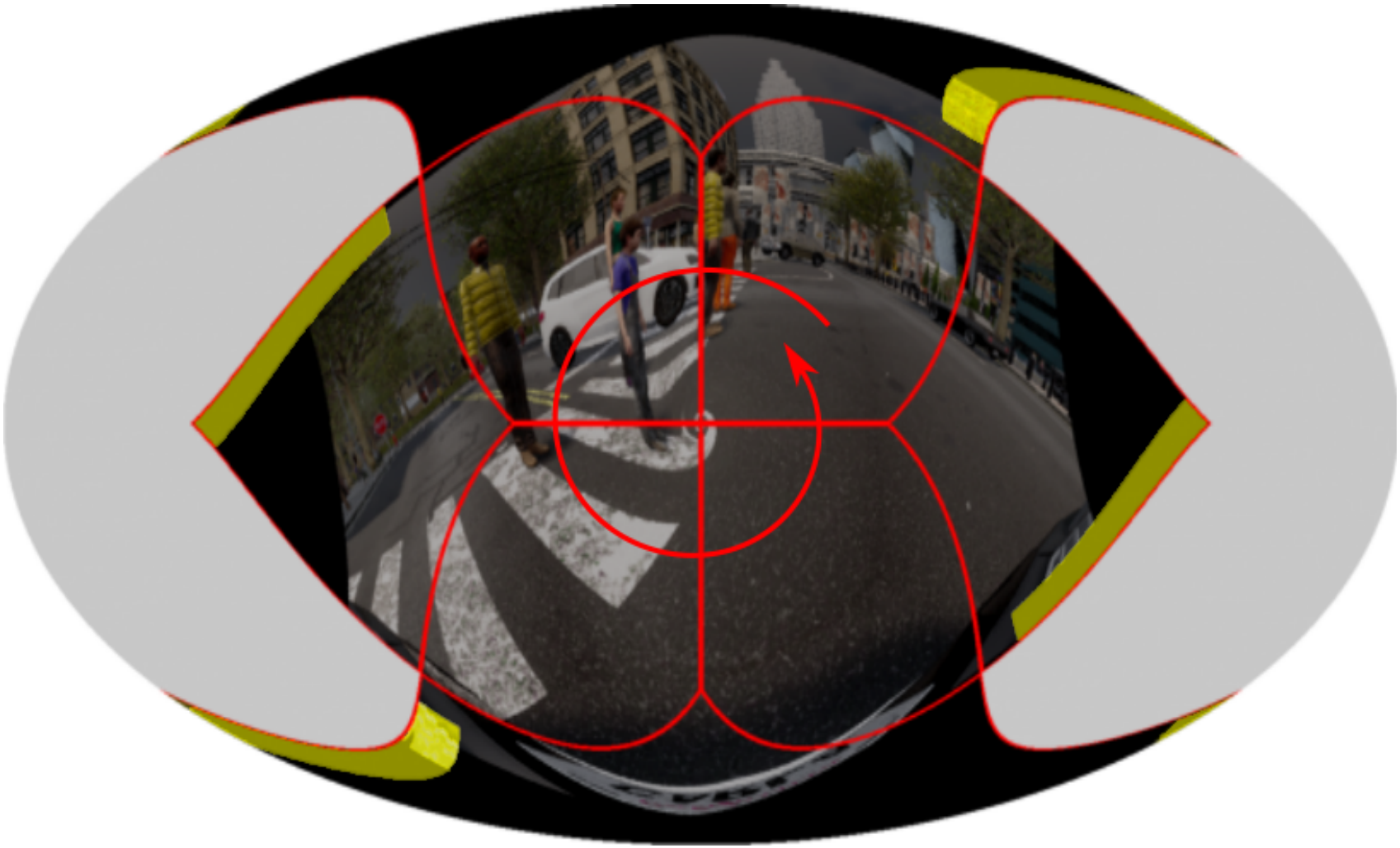


HEAL-SWIN



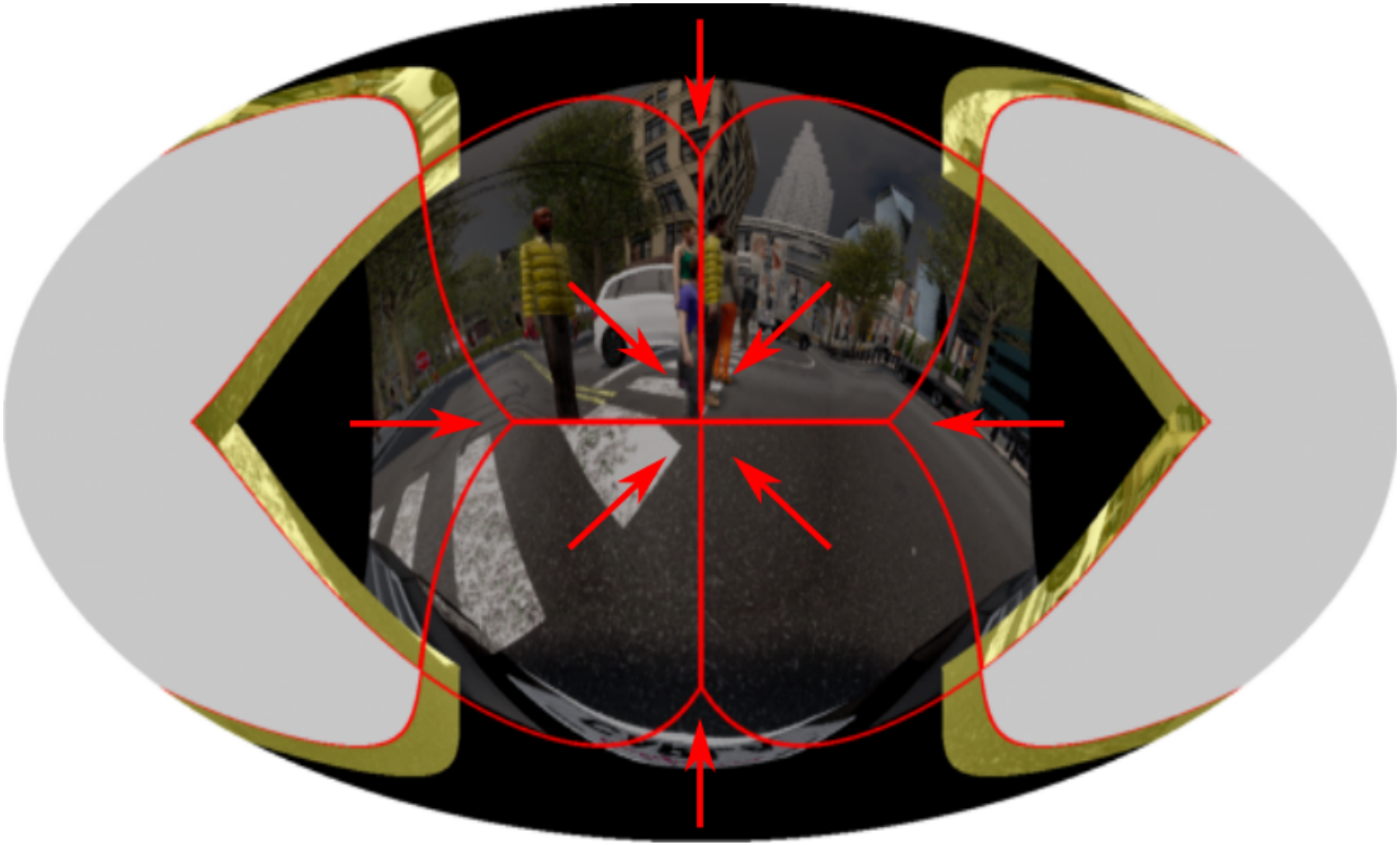
SWIN

HEAL-SWIN spiral window shift



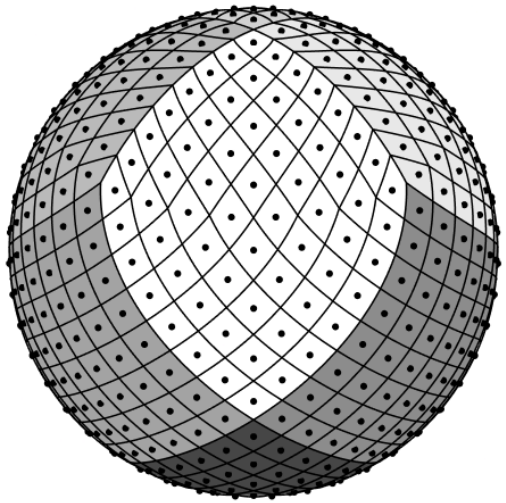
Spiral shifting.

HEAL-SWIN nested window shift

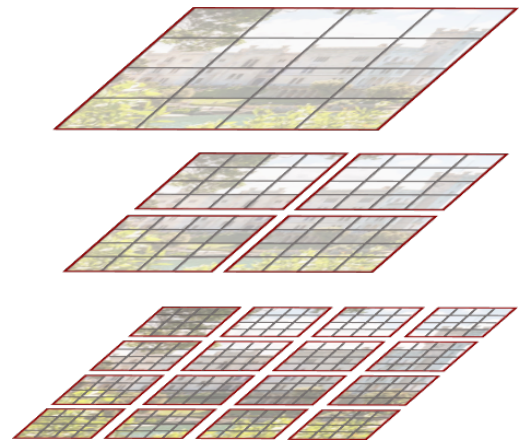


Nested shifting.

Equivalent architecture



HEAL-SWIN



SWIN

Allows for equivalent architecture with identical number of parameters.

Experiments and results

Datasets

- WoodScape
  - Semantic segmentation

Datasets

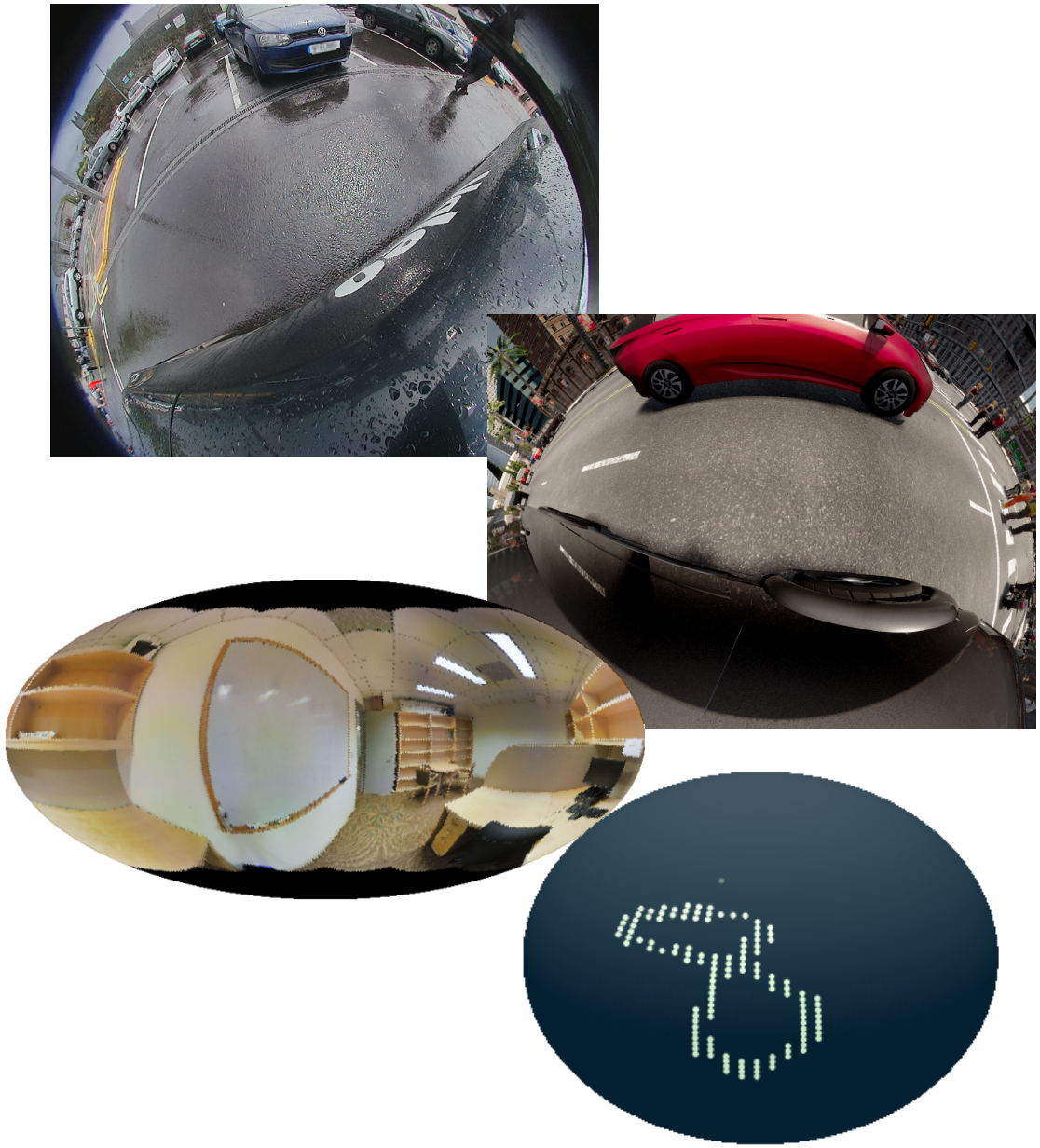
- WoodScape
  - Semantic segmentation
- SynWoodScape
  - Semantic segmentation
  - Depth estimation

Datasets

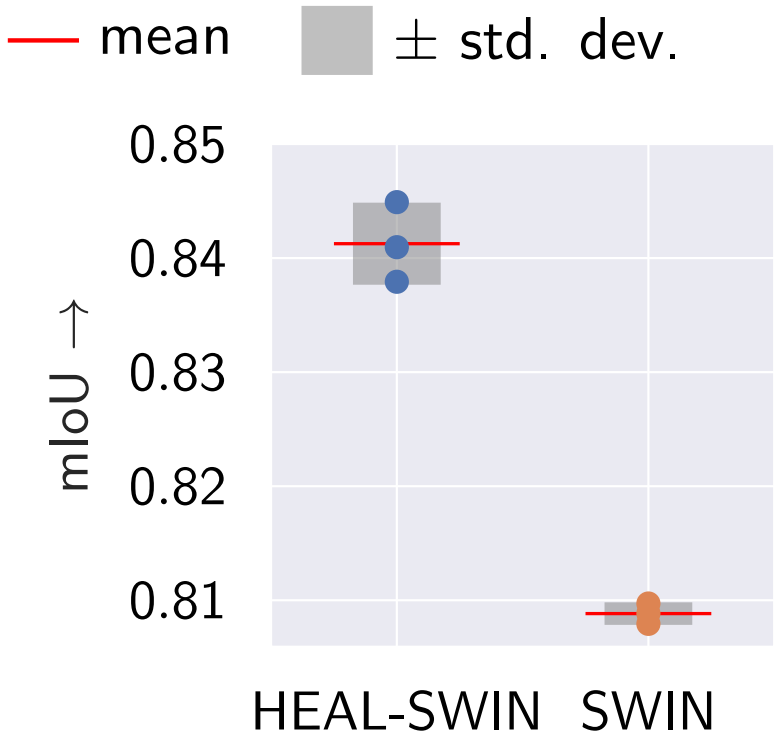
- WoodScape
  - Semantic segmentation
- SynWoodScape
  - Semantic segmentation
  - Depth estimation
- Stanford 2D3D-S
  - Semantic segmentation

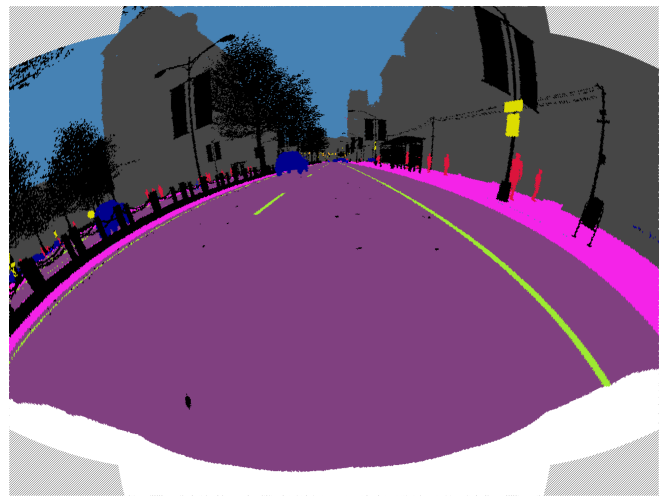
Datasets

- WoodScape
  - Semantic segmentation
- SynWoodScape
  - Semantic segmentation
  - Depth estimation
- Stanford 2D3D-S
  - Semantic segmentation
- MNIST
  - Classification

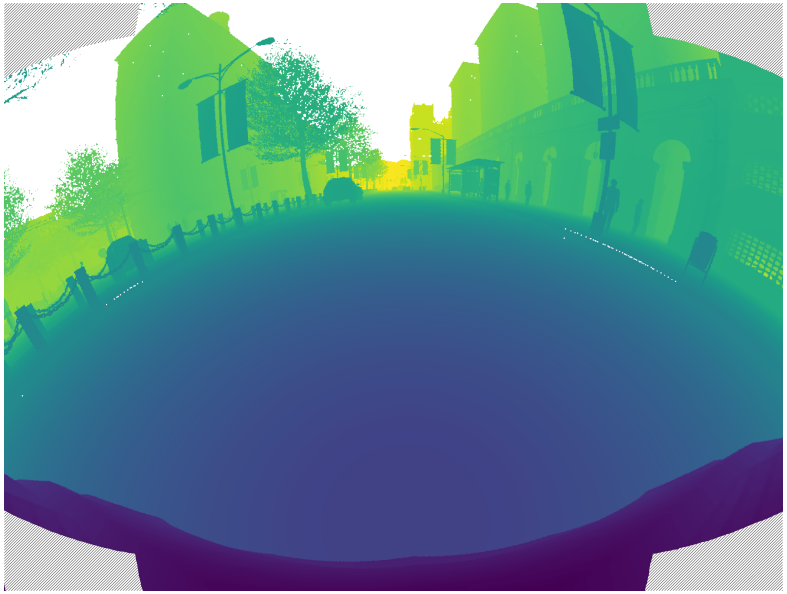


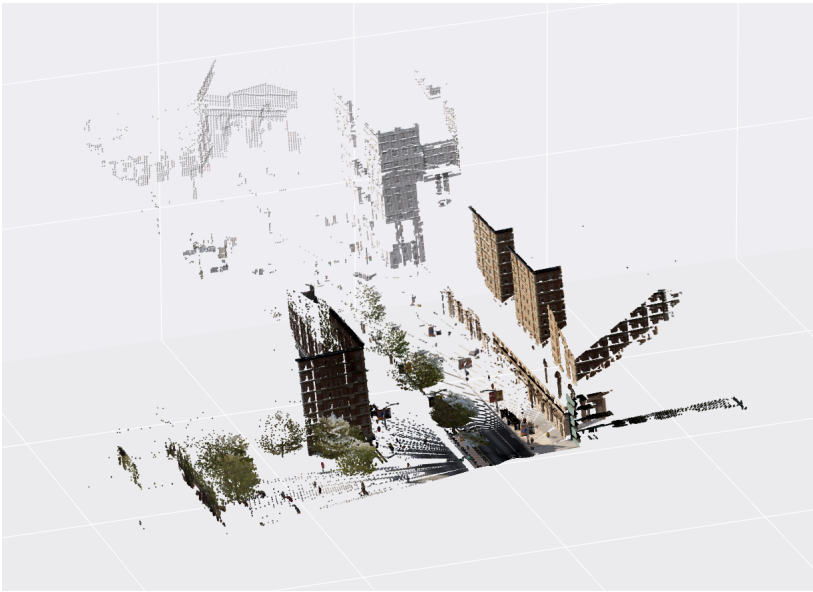
Results: Semantic segmentation on SynWoodScape



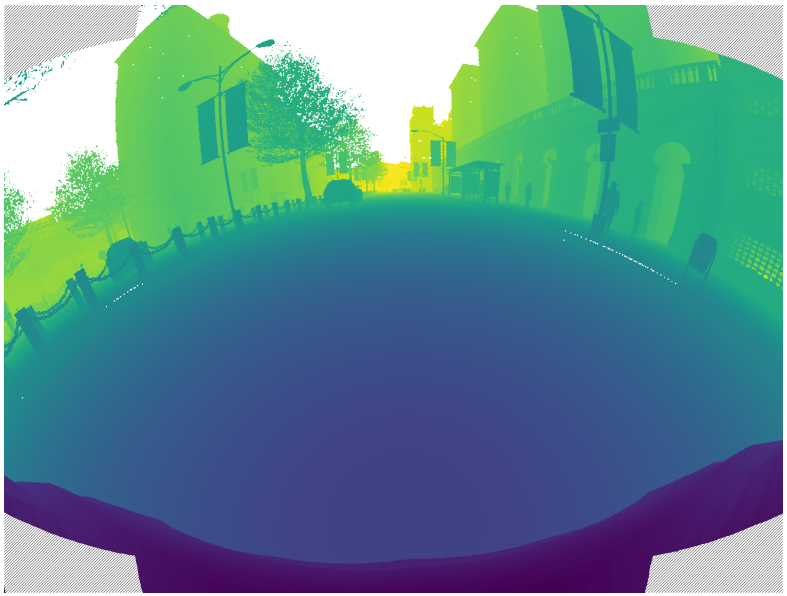


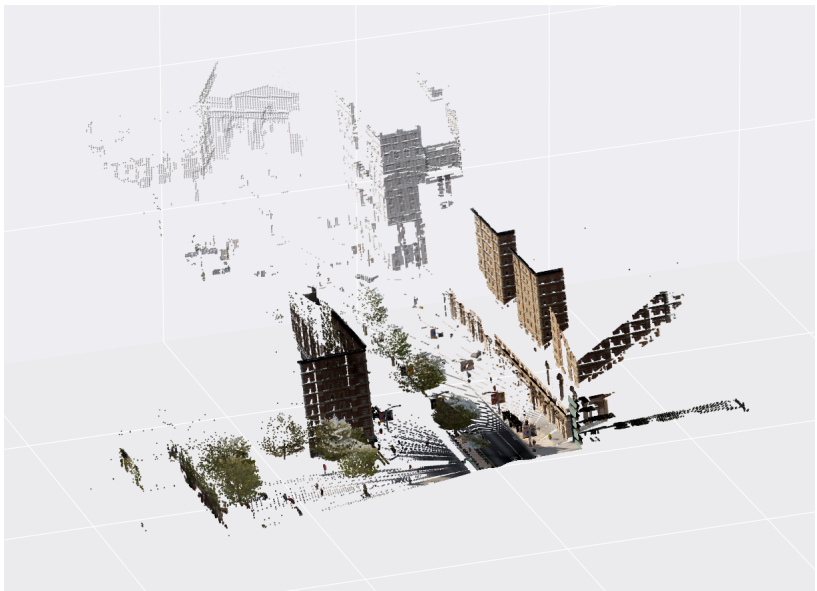
Experiment: Depth estimation





Experiment: Depth estimation

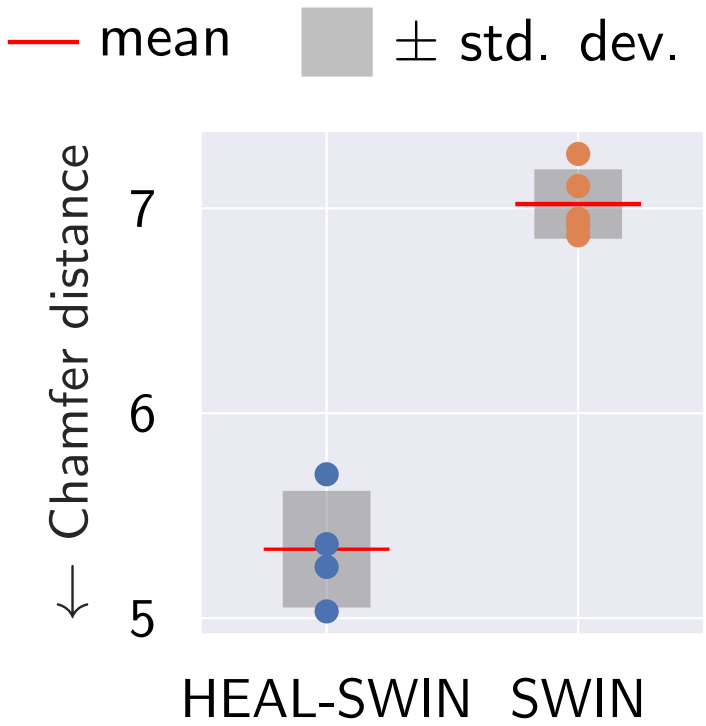


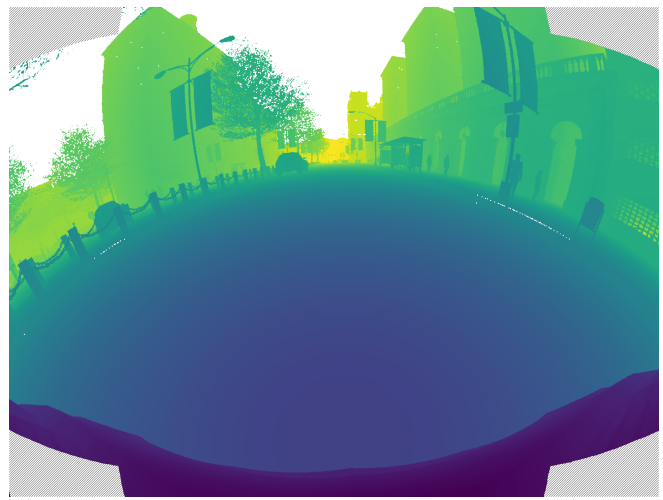


Chamfer distance:

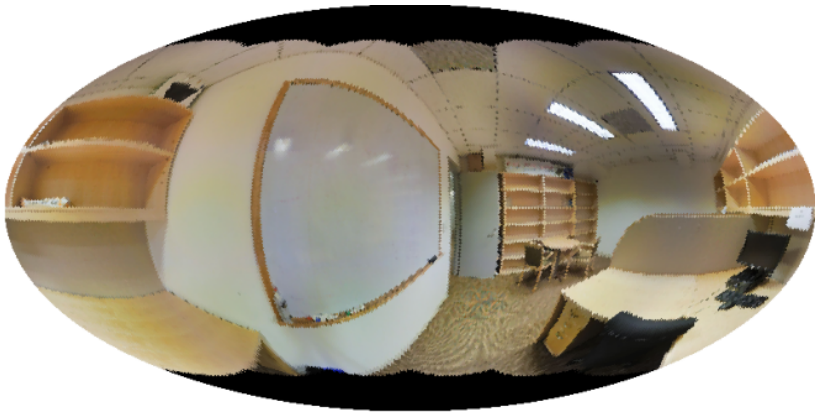
$$\mathrm{CD}(P_{\mathrm{pred}}, P_{\mathrm{gt}}) = \frac{1}{|P_{\mathrm{pred}}|} \sum_{p \in P_{\mathrm{pred}}} \min_{p' \in P_{\mathrm{gt}}} d(p, p')^2 + \frac{1}{|P_{\mathrm{gt}}|} \sum_{p \in P_{\mathrm{gt}}} \min_{p' \in P_{\mathrm{pred}}} d(p, p')^2$$

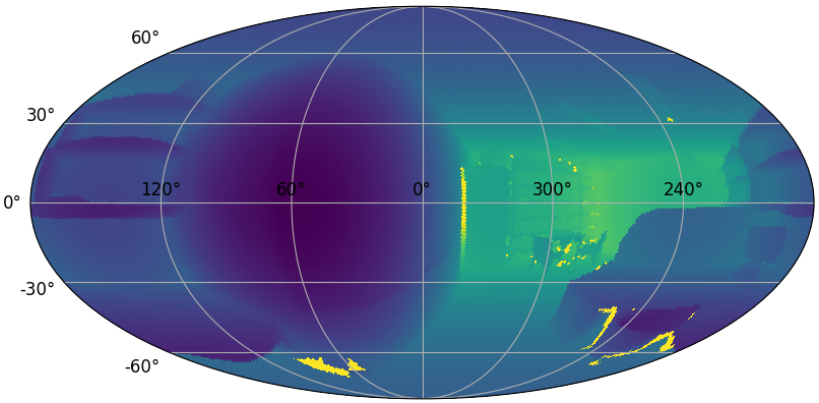
Results: Depth estimation on SynWoodScape

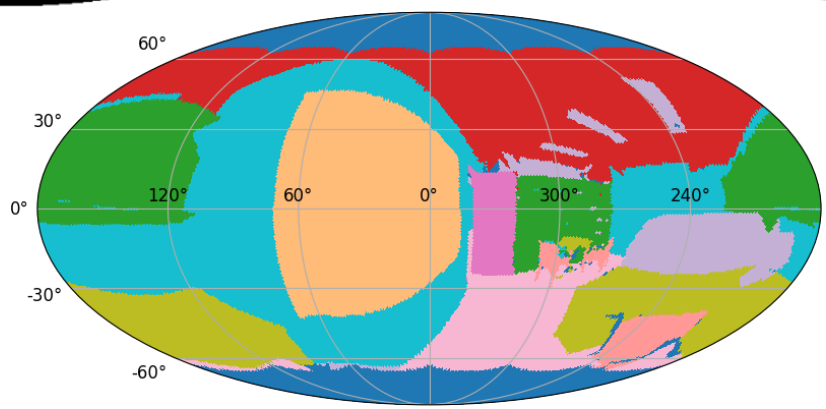




Experiment: Semantic segmentation on Stanford 2D3D-S

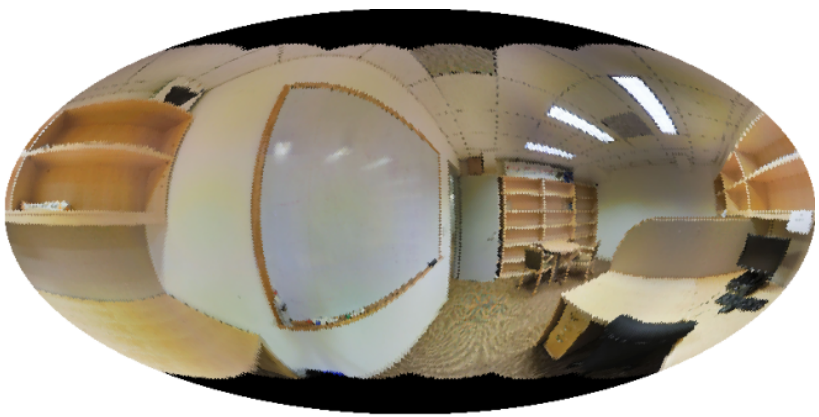


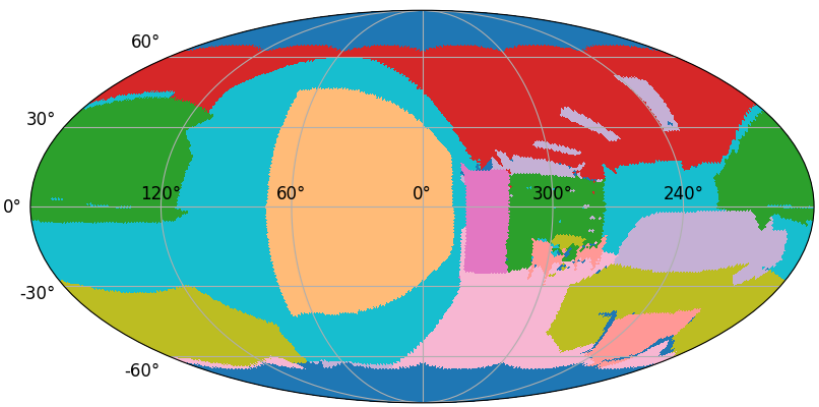




Results: Semantic segmentation on Stanford 2D3D-S

Model	mIoU	mAcc
Gauge CNN (Cohen et al., 2019)	39.4	55.9
UGSCNN (Jiang et al., 2019)	38.8	54.7
HexRUNet (Zhang et al., 2019)	43.3	58.6
SphCNN (Esteves et al., 2018; 2020)	40.2	52.8
Spin-SphCNN (Esteves et al., 2020)	41.9	55.6
HEAL-SWIN (Ours)	**44.3**	**61.9**





Outlook

- Combine HEAL-SWIN (or other HEALPix based models) with $\mathrm{SO}(3)$ equivariance
- Model weather behaviour using HEALPix grid
- Examine the foundational mathematical structures of transformers in search of a unifying framework for equivariant transformers

Thanks for your attention!

Questions?

Link to our paper:

Accepted as poster to CVPR 2024.