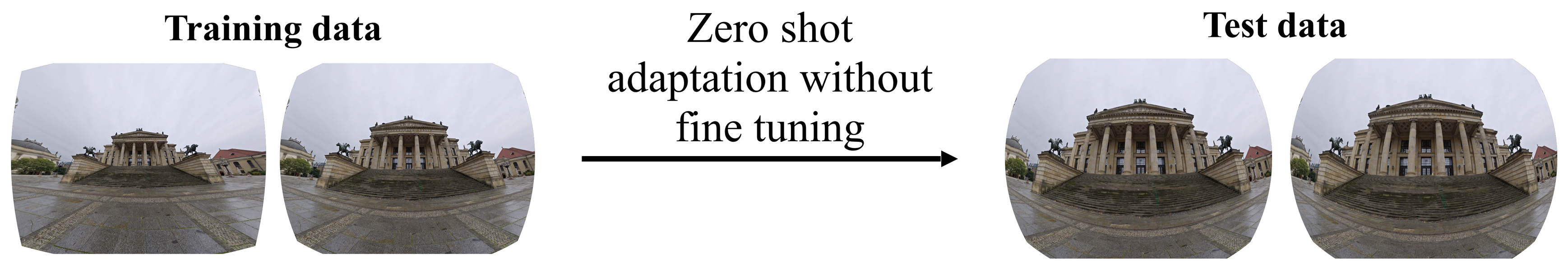
DarSwin: Distortion Aware Radial Swin Transformer

Akshaya Athwale[1], Arman Afrasiyabi[3], Justin Lagüe[1], Ichrak Shili[1], Ola Ahmad[2], Jean-François Lalonde[1]

Université Laval[1], Thales Digital Solution[2], Yale University[3]

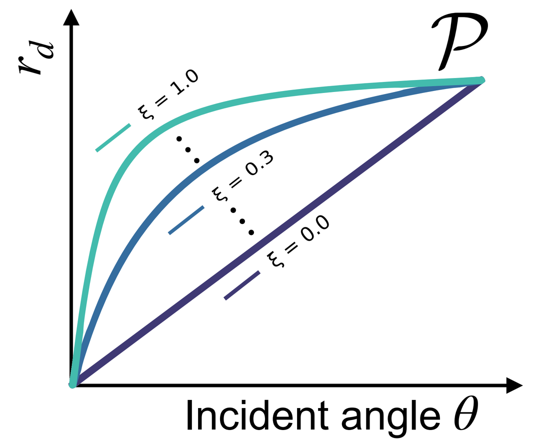
Motivation

- Wide-angle lenses produce significant distortion. Lens distortion profile depends on the type of lens.
- An approach naively trained for a specific lens overfits to its specific distortion and does not generalize well when tested on another lens.
- Rectification of distortion leads to loss of field of view and creates artifacts.
- We present DarSwin, which embeds the physical characteristics of lenses and performs zero-shot adaptation to unknown and out-of-distribution lens profiles without fine-tuning.



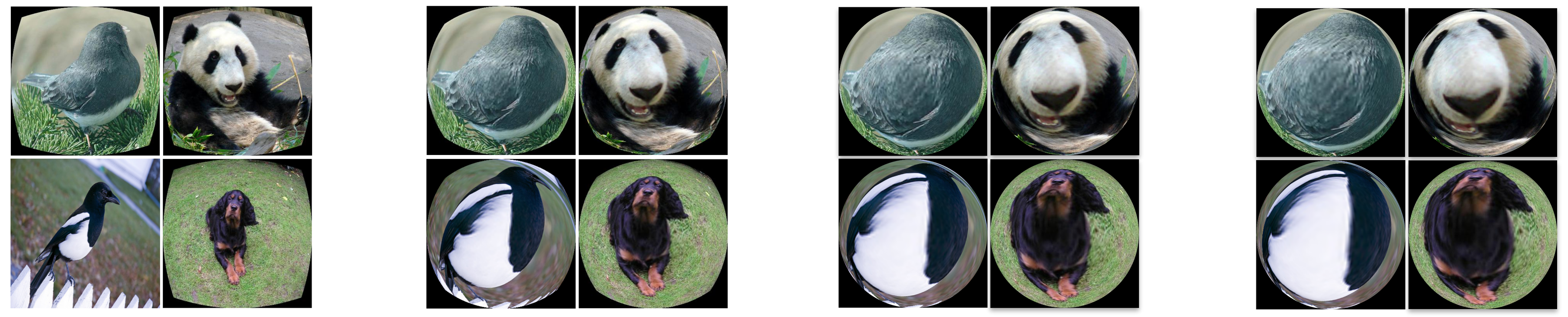
Synthetically distorted ImageNet with radial distortion

- Spherical distortion : one parameter $\xi \in [0,1]$
- Projection function : $\mathcal{P} = F(r_d, \theta)$
- Distance to images center : $r_d = f \dfrac{\sin(\theta)}{\cos(\theta) + \xi}$



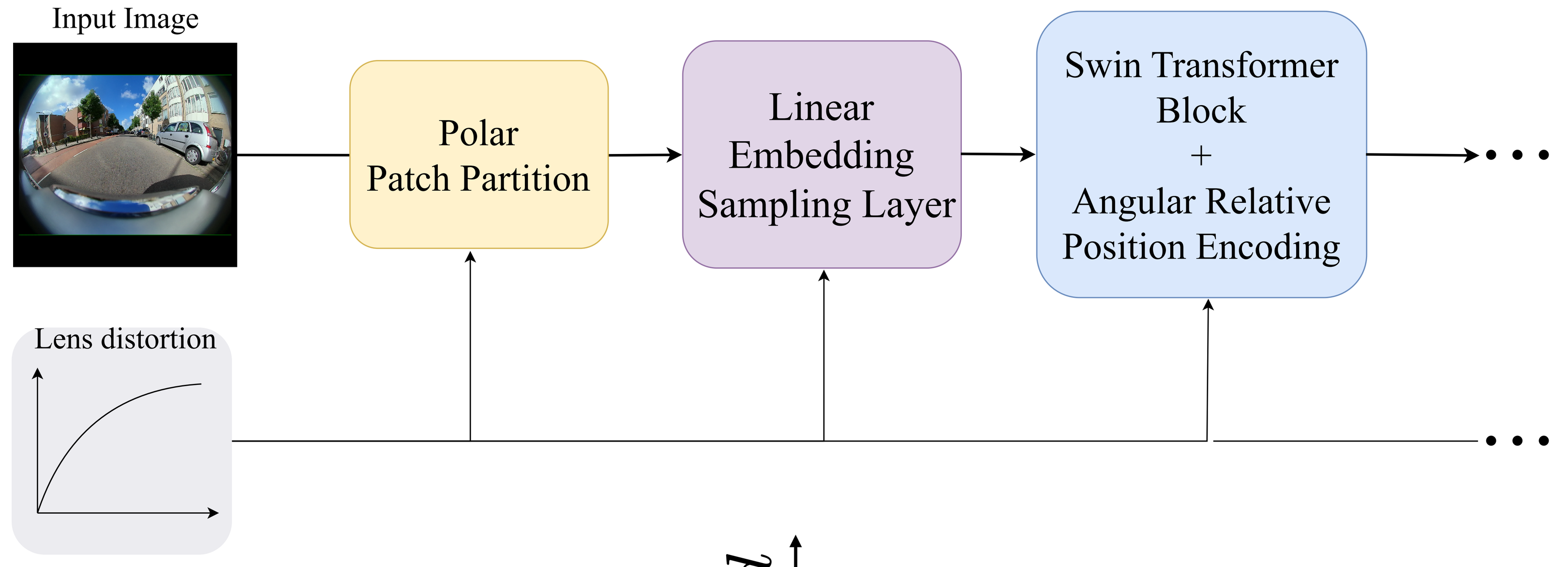
Training sets based on different levels of distortion

Very Low	Low	Medium	High
$\xi \in [0.0, 0.05]$	$\xi \in [0.2, 0.35]$	$\xi \in [0.5, 0.70]$	$\xi \in [0.8, 0.95]$



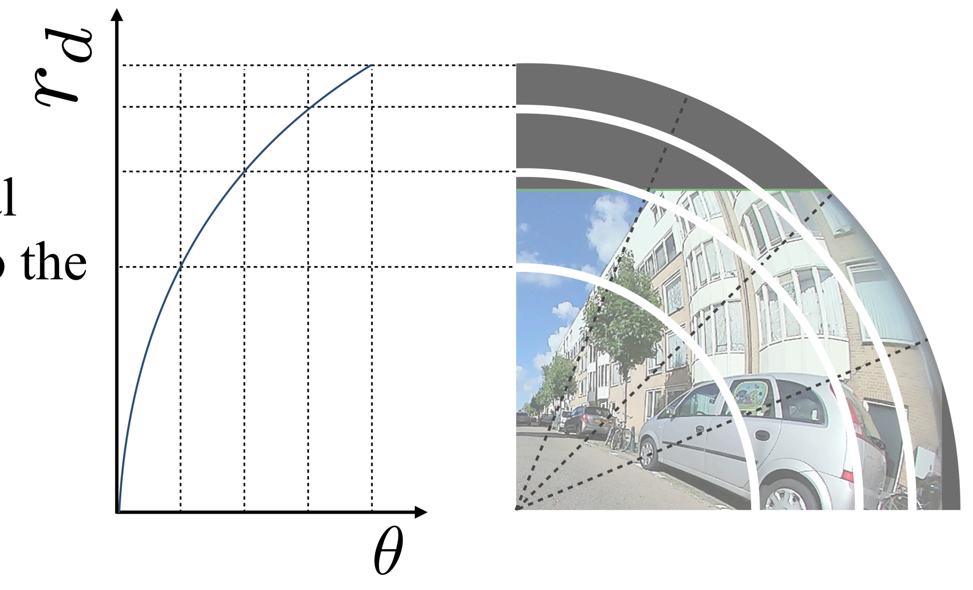
Evaluation on generalization to all $\xi \in [0,1]$

Method

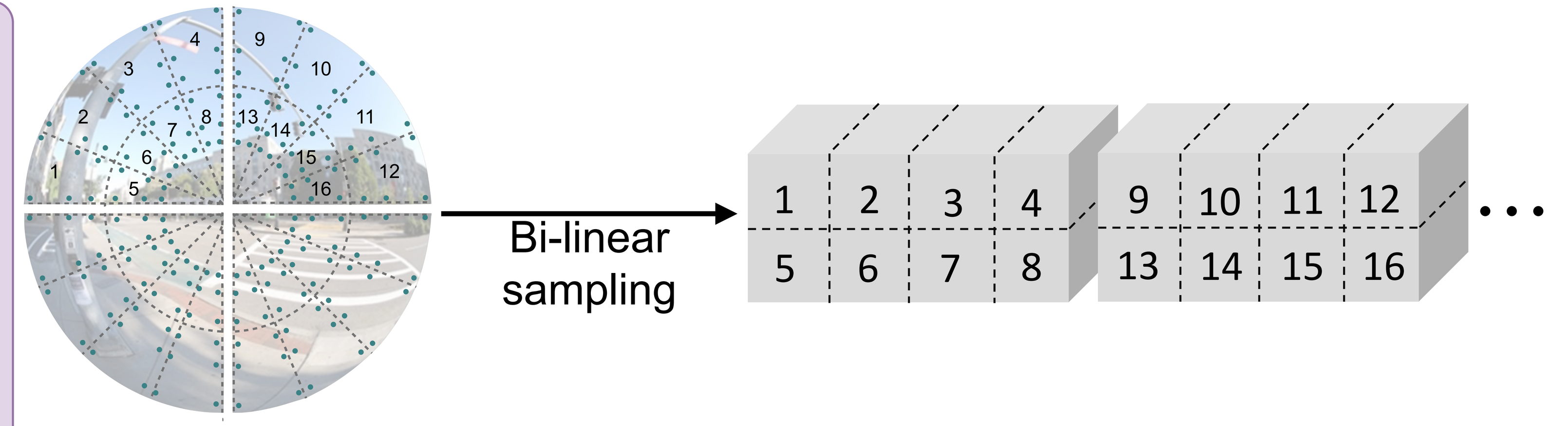


Polar Patch Partition

Key idea: embedding physical characteristics of the lens into the network through distortion-aware radial partition.

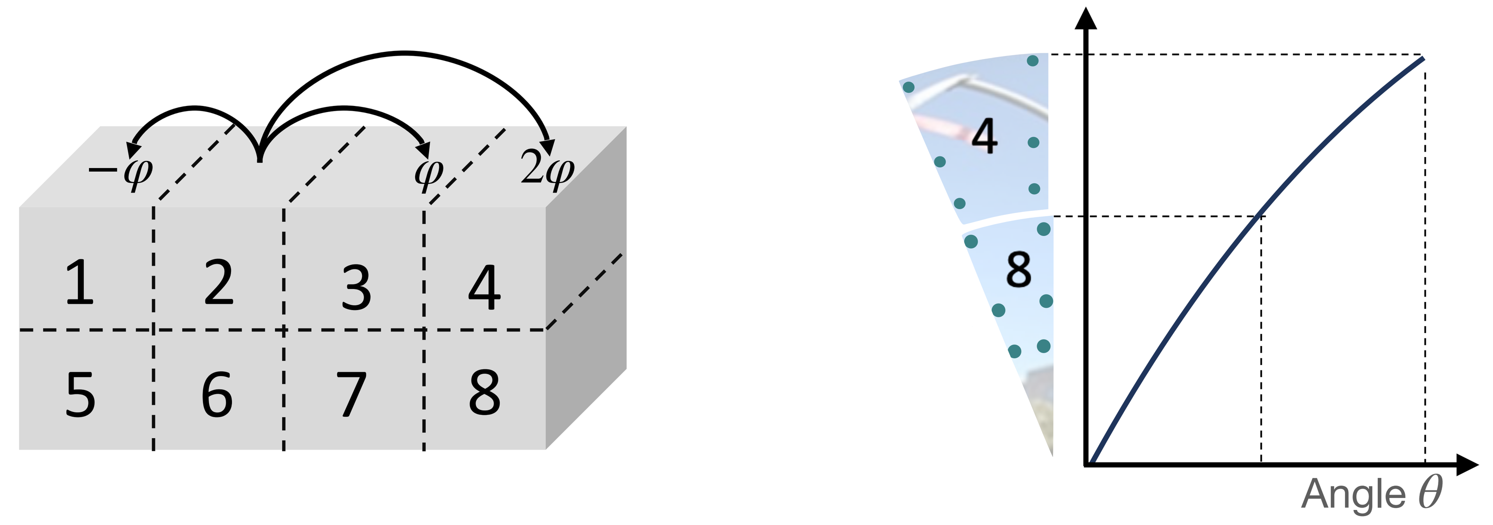


Linear Embedding Sampling Layer



This polar format is used to linearly embed each patch using an MLP layer.

Swin Transformer block + Angular Relative Positional Encoding



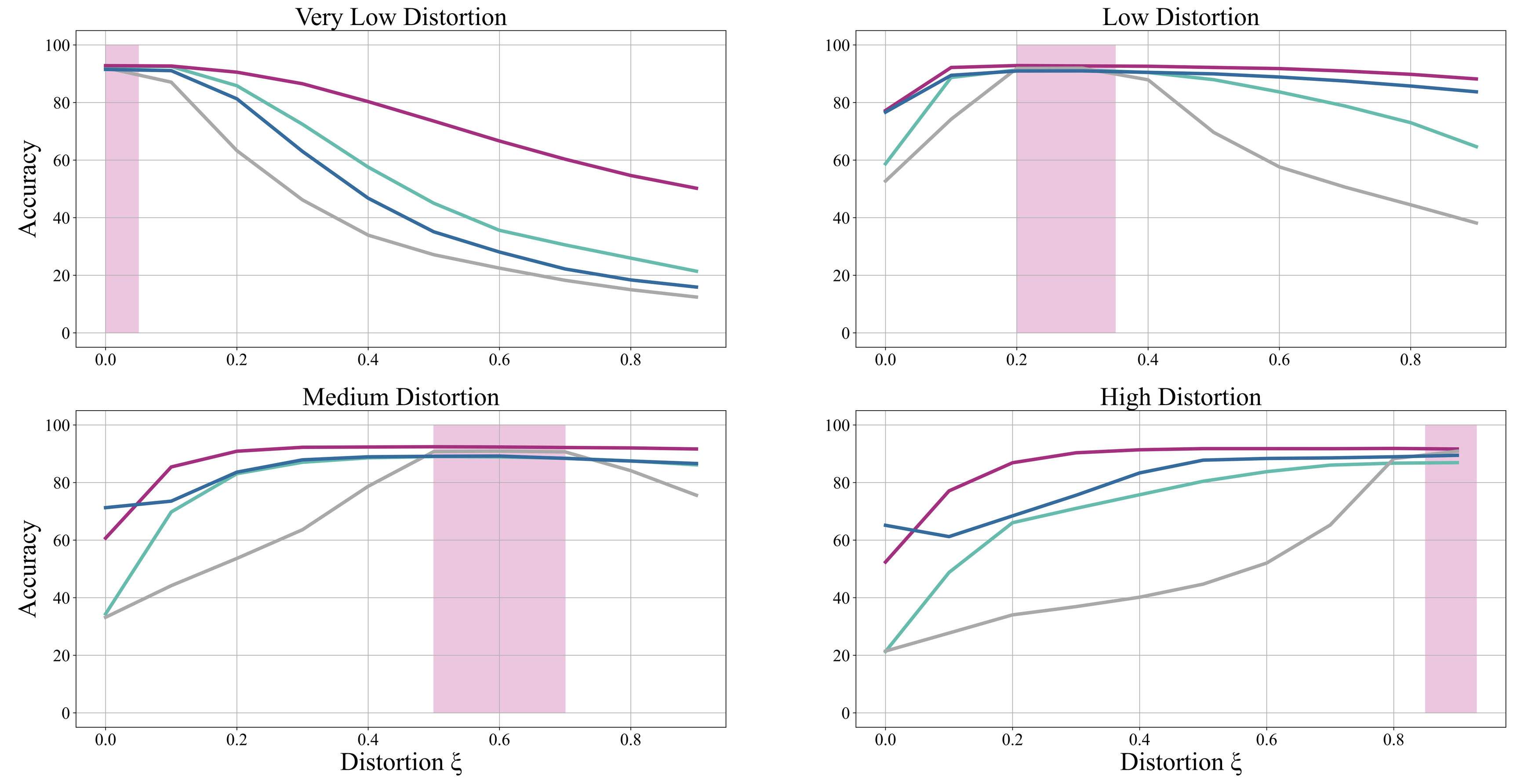
$B_\varphi = a_{\Delta\varphi}\sin(\Delta\varphi) + b_{\Delta\varphi}\cos(\Delta\varphi)$ $\quad B_\theta = a_{\Delta\theta}\sin(\Delta\theta) + b_{\Delta\theta}\cos(\Delta\theta)$

$Att(Q,K,V) = Softmax(QK^T/\sqrt{d} + B_\theta + B_\varphi)V$

We parameterize a smaller-sized bias matrix $\hat{B}_\theta \in \mathbb{R}^{(2M_\theta - 1)\times 2}$ and $\hat{B}_\varphi \in \mathbb{R}^{(2M_\varphi - 1)\times 2}$ and values in a_* and b_* are taken from $\hat{B}_\theta$ and $\hat{B}_\varphi$.

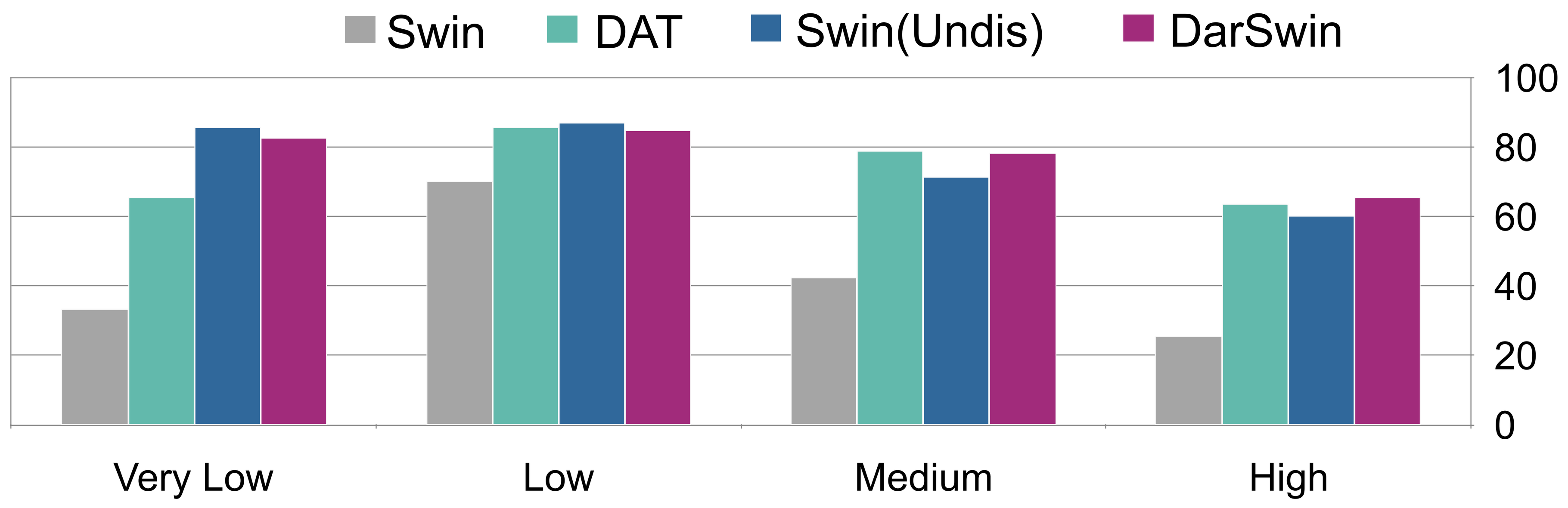
Results

Generalization to other distortions out of training distribution

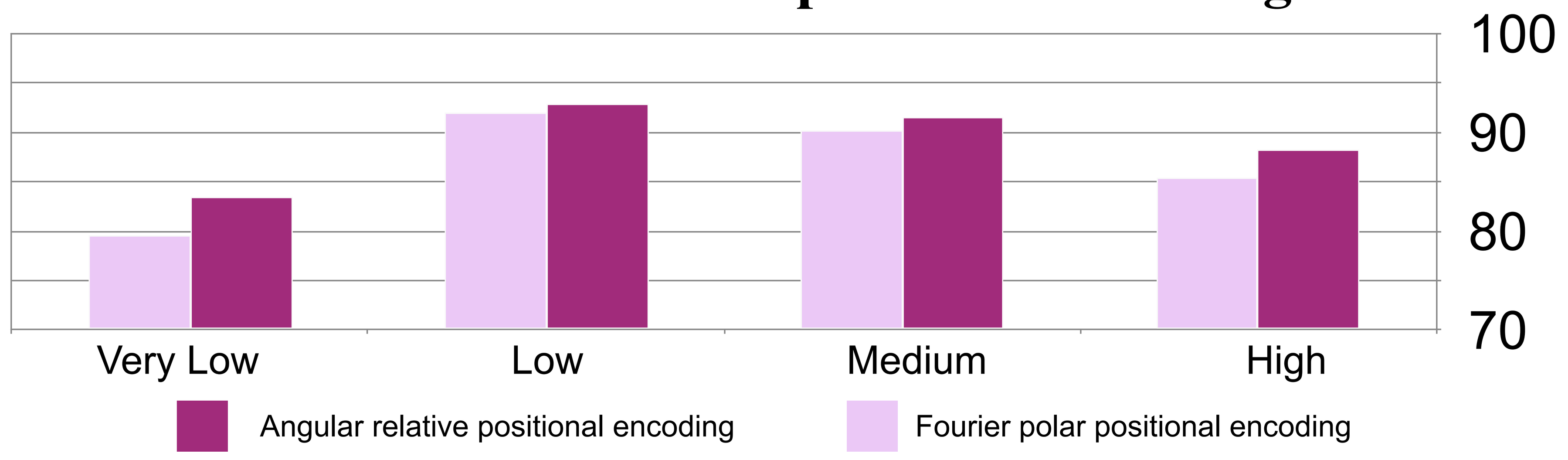


- Training distortion distribution.
- Swin: Swin transformer trained on distorted images.
- DAT: Deformable attention transformer trained on distorted images.
- Swin(undis): Swin transformer trained on images undistorted using true un-distortion from spherical projection.
- DarSwin: Our method trained on distorted images.

Generalization to projection models (Polynomial projection)



DarSwin trained with different positional encodings



ICCV23