Xray$_{to}$3DShape Benchmark

Benchmarking Encoder-Decoder Architectures for Biplanar X-ray to 3D Shape Reconstruction

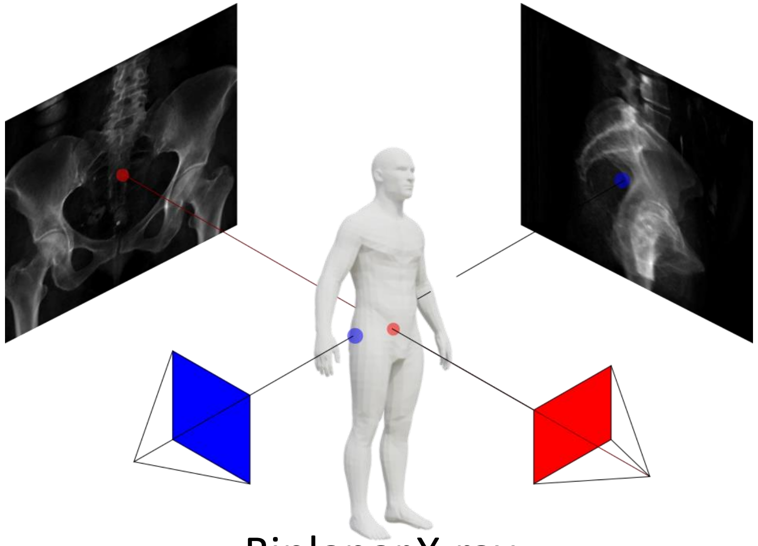
Mahesh Shakya and Bishesh Khanal

TransfOrming Global health with AI (**TOGAI**) Lab,
Nepal Applied Mathematics and Informatics Institute for research (**NAAMII**)

NeurIPS 2023 Datasets and Benchmarks Track

NAAMII

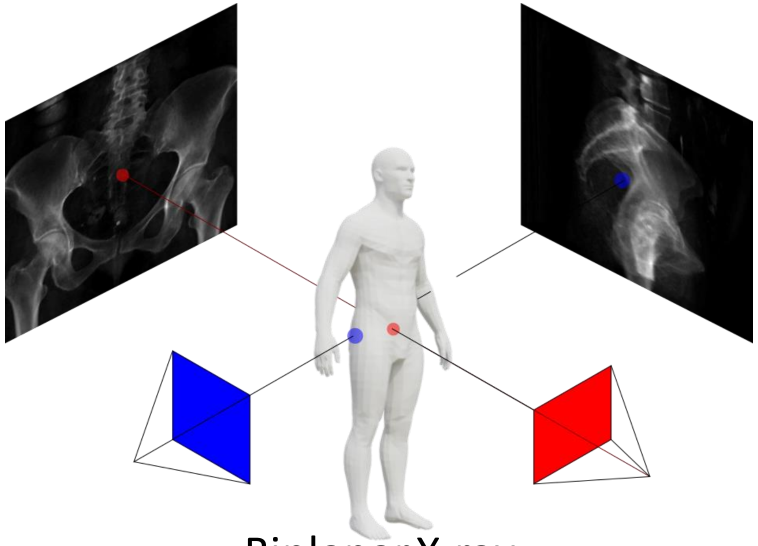
Biplanar Xray to 3D Reconstruction



Biplanar X-ray

Reconstructed
3D Shape

Biplanar Xray to 3D Reconstruction



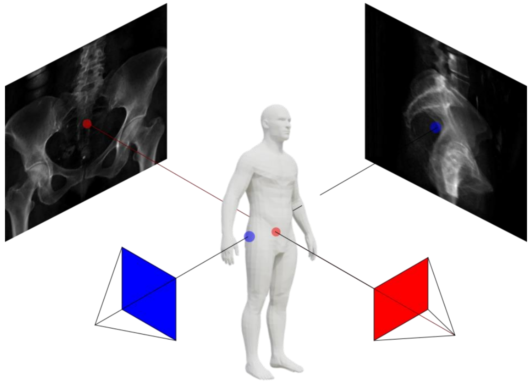
Biplanar X-ray

Reconstructed
3D Shape

Augments low cost, low radiation X-ray device with CT-like 3D visualization

Improved diagnosis, surgery planning and navigation, better biomarkers

Biplanar Xray to 3D Reconstruction



Augments low cost, low radiation X-ray device with CT-like 3D visualization

Improved diagnosis, surgery planning and navigation, better biomarkers

Limitations of existing works

Validation

- single **private dataset**
- limited baseline comparison

Evaluation

only image-based metrics

Analysis

only aggregated metrics

Procedure Reference→	Previous Approaches			
	UNet Kasten[2020]	Transvert Bayat[2020]	TL-Embedding Shiode[20212]	1DConcat Chen[2019]
Dataset	Private	Mixed	Private	Private
Anatomy of Interest	Knee	Vertebra	Wrist	Vertebra
Input views	AP & LAT	AP & LAT	AP	AP & LAT
Input Resolution(mm)	1.0	1.5	0.4	1.5
Input Size	128^2	64^2	500×625	64^2
Training Samples	188	~10k	147	90
Test Samples	20	~2k	26	10
Supervised Loss	weighted-CE	L1	CE	L2
Adversarial Loss	✗	✓	✗	✗
Reprojection Loss	✓	✗	✗	✗
AP/LAT View-Fusion	Input-level	Feature-level	AP view only	Feature-level
Surface Error(mm)				
↳ avg	1.778	✗	1.05 - 1.45	✗
↳ max	✗	5.11	✗	✗
Dice Score	90.7	95.5	✗	74.0

Limitations of existing works

Validation
- single private dataset
- limited baseline comparison

Evaluation
only **image-based metrics**

Analysis
only **aggregated metrics**

Procedure Reference→	Previous Approaches			
	UNet Kasten[2020]	Transvert Bayat[2020]	TL-Embedding Shiode[20212]	1DConcat Chen[2019]
Dataset	Private	Mixed	Private	Private
Anatomy of Interest	Knee	Vertebra	Wrist	Vertebra
Input views	AP & LAT	AP & LAT	AP	AP & LAT
Input Resolution(mm)	1.0	1.5	0.4	1.5
Input Size	128^2	64^2	500×625	64^2
Training Samples	188	~10k	147	90
Test Samples	20	~2k	26	10
Supervised Loss	weighted-CE	L1	CE	L2
Adversarial Loss	✗	✓	✗	✗
Reprojection Loss	✓	✗	✗	✗
AP/LAT View-Fusion	Input-level	Feature-level	AP view only	Feature-level
Surface Error(mm)				
↳ avg	1.778	✗	1.05 - 1.45	✗
↳ max	✗	5.11	✗	✗
Dice Score	90.7	95.5	✗	74.0

Key Contributions of Xray-to-3D-benchmark

Validation: First comprehensive benchmark

Evaluation: Elaborate Evaluation including Clinical Tasks and Metrics

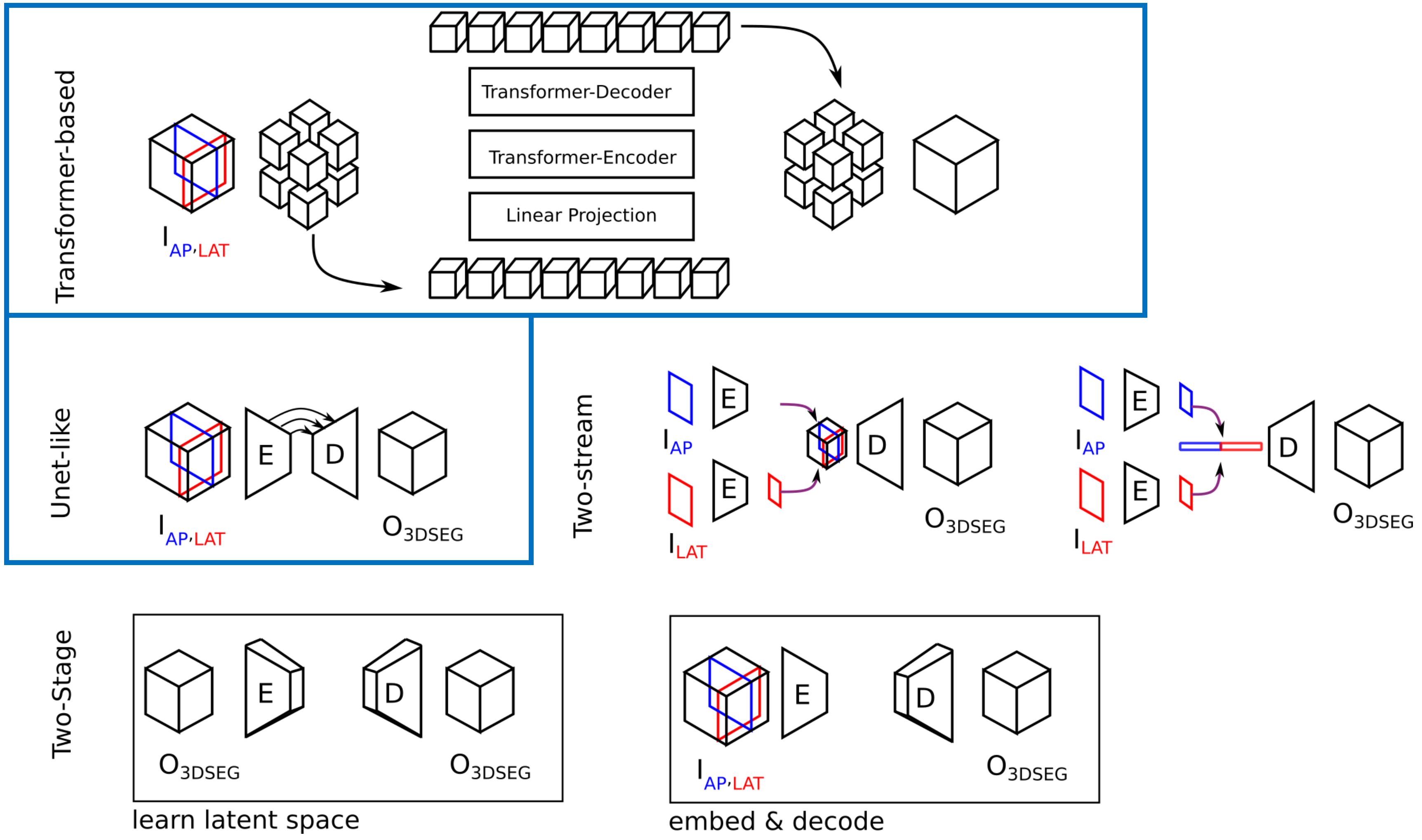
Analysis: Disaggregated reporting

Reproducible and accessible benchmark toolkit

4 anatomies, 6 Benchmarked Datasets

CT Segmentation Datasets	**Anatomy**	**Preprocessing**			
		Reject partial bones	Reject anisotropic samples	Extract ROI	Segment u/ Pretrained model
CTSpine1K		✓	✓	✓	
CTPelvic1K		✓	✓	✓	
TotalSegmentator	  	✓		✓	
LIDC-IDRI		✓		✓	✓
VerSe2019		✓		✓	
RSNA Cervical Fracture		✓		✓	

8 Benchmarked Encoder-Decoder Architecture



Transformer-based

- SwinUNETR (2022), UNETR (2022)

UNet-like

- Attention-UNet (2018), UNet (2015)

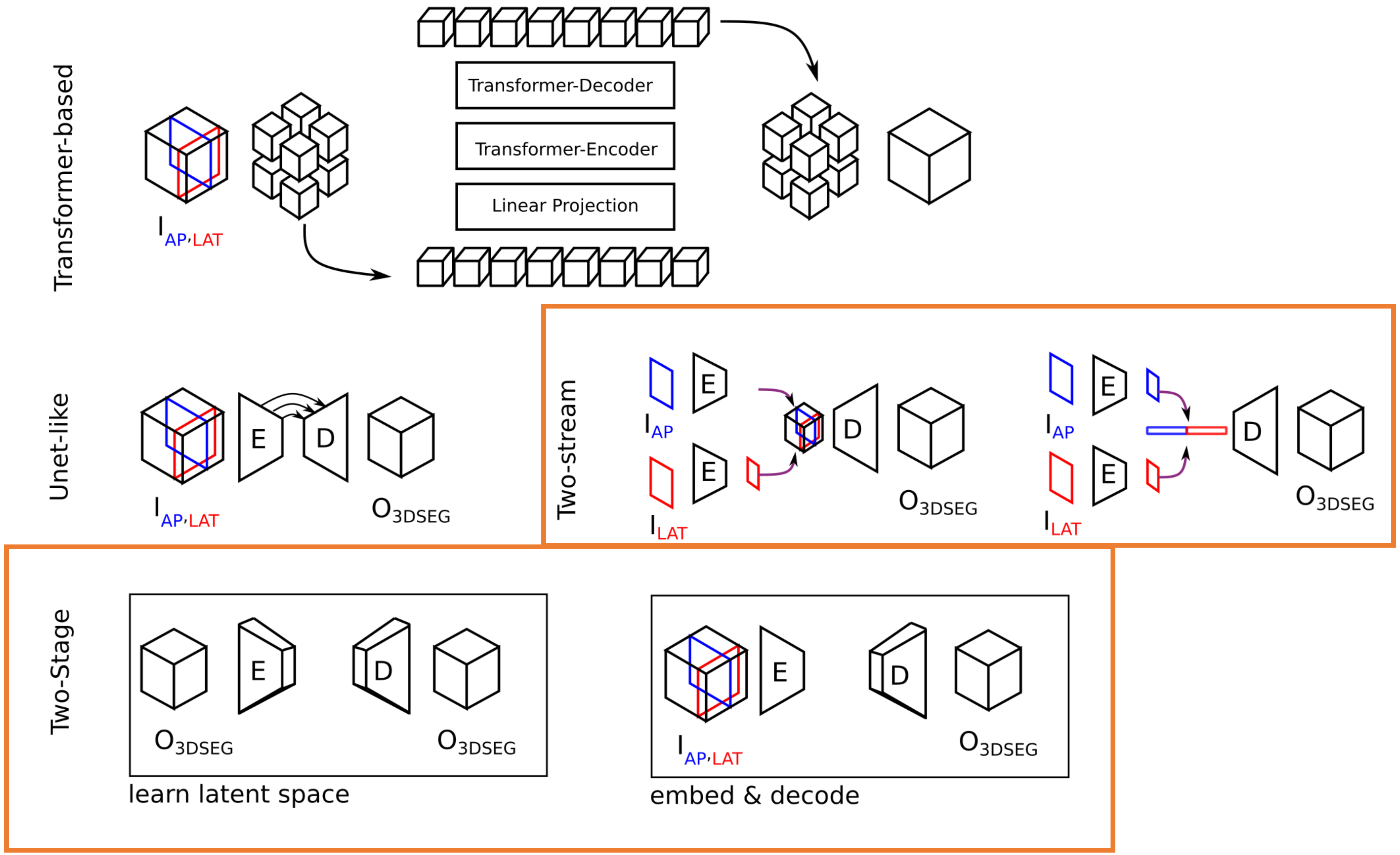
Two-Stage Enc-Decoder

- TL-Embedding (2016)

Two-stream Enc-Decoder

- 1D-Concat, 2D-Concat (2020), MultiScaleConcat (2019)

8 Benchmarked Encoder-Decoder Architecture



Transformer-based

- SwinUNETR (2022), UNETR (2022)

UNet-like

- Attention-UNet (2018), UNet (2015)

Two-Stage Enc-Decoder

- TL-Embedding (2016)

Two-stream Enc-Decoder

- 1D-Concat, 2D-Concat (2020), MultiScaleConcat (2019)

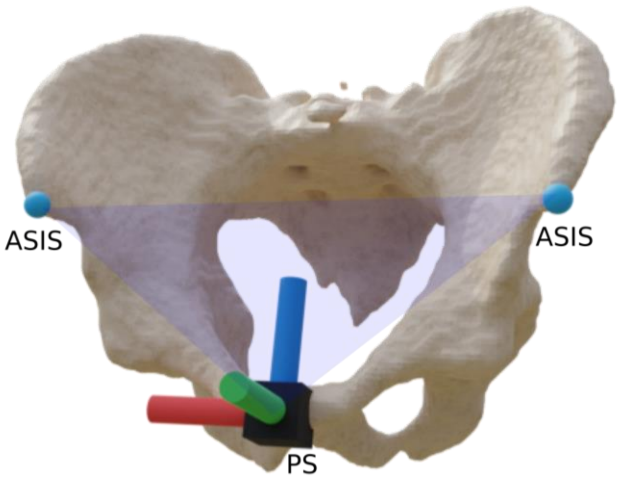
Clinically relevant metrics

Reconstructed
3D Shape

Clinically relevant metrics

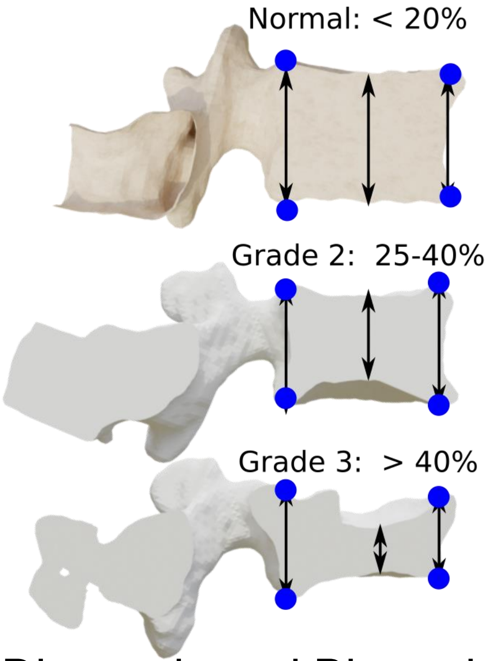
Reconstructed
3D Shape

Applications / Downstream tasks



Patient-specific Modelling

Pelvic Coordinate System : for standardized reporting of joint biomechanics

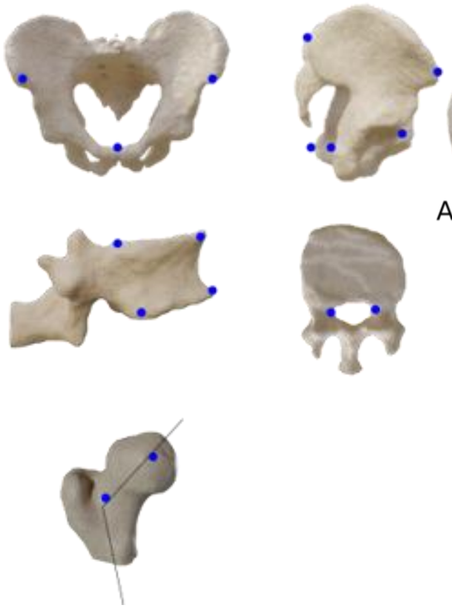


Diagnosis and Biomarker

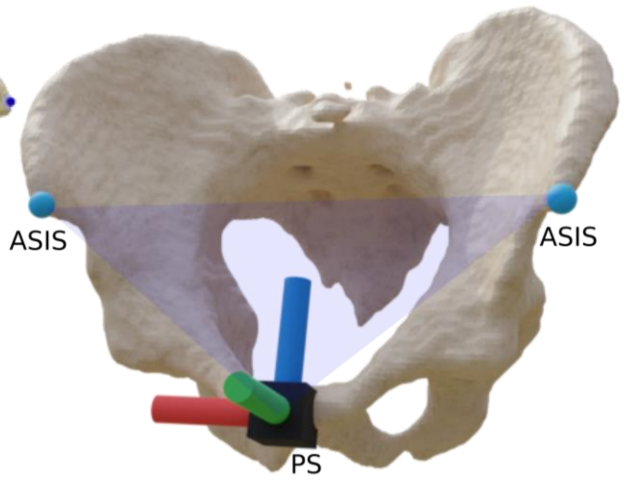
Compression Fracture Grading: relative reduction of vertebra height

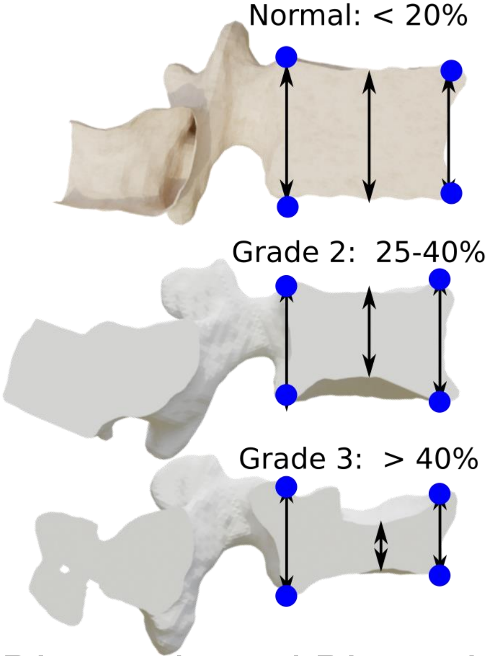
Clinically relevant metrics

Reconstructed
3D Shape



Applications / Downstream tasks





Patient-specific Modelling

Pelvic Coordinate System : for standardized reporting of joint biomechanics

Diagnosis and Biomarker

Compression Fracture Grading: relative reduction of vertebra height

Result Highlights

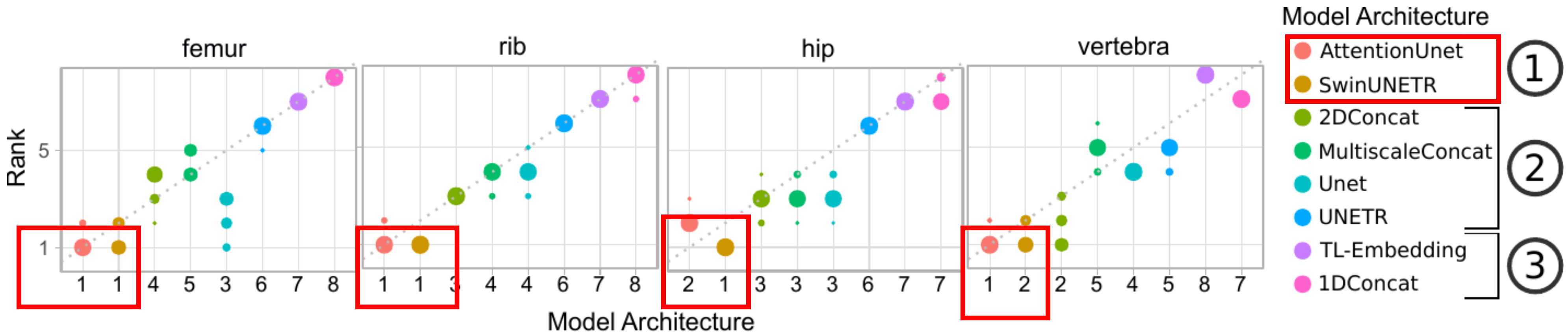
Benchmarking : Which architecture performs best?
SwinUNETR and AttUnet perform well across multiple datasets and anatomies

Performance gaps for clinical deployment
- Reduced performance on clinically relevant minority subgroups
- Reduced performance on task-specific domain shifts such as fractures, misaligned x-rays

Clinically relevant metrics
- Dice Score for model evaluation at dataset-level does not do justice at Patient-level

SwinUnetr and AttUNet perform well across datasets

Dataset (#train/#test) (vol size) (voxel resolution)	Method Reference	#Param	Dice(%)↑	HD95(mm)↓	ASD(mm)↓	NSD@1.5mm↑
Aggregate	SwinUNETR	62.2M	**79.27**	3.65	0.86	0.68
	AttentionUnet	1.5M	78.92	**3.07**	**0.84**	**0.69**
	TwoDPermuteConcat	1.2M	78.08	3.33	0.91	0.67
	UNet	1.2M	77.27	3.49	1.00	0.66
	MultiScale2DPermuteConcat	3.5M	77.09	4.16	0.96	0.65
	UNETR	96.2M	74.20	4.27	1.14	0.62
	TLPredictor	6.6M	69.53	4.70	1.43	0.54
	OneDConcat	40.6M	69.16	7.07	1.53	0.53

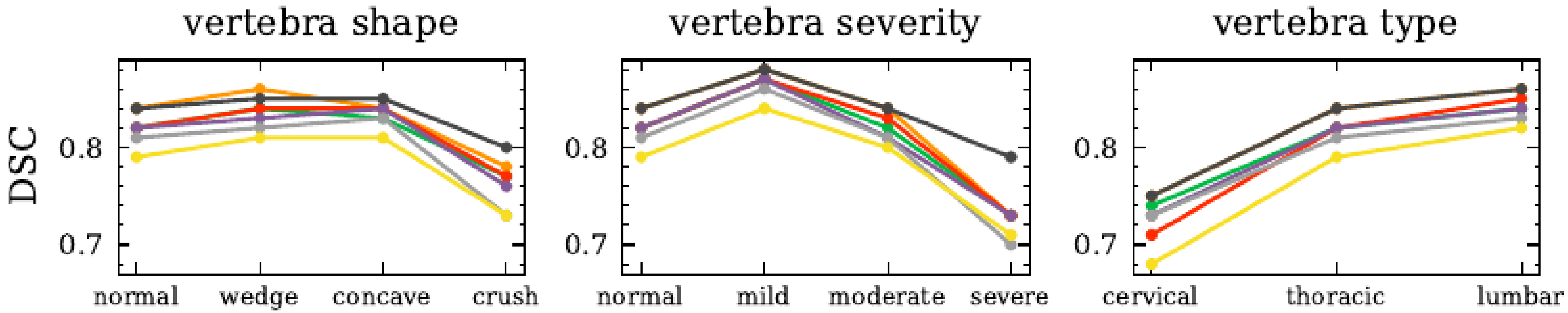


Performance gaps for clinical deployment

Discrepancy in performance on clinically relevant minority subgroups

Reduced performance for Image & Population Domain shifts hinder clinical translation

Misaligned X-ray views can slightly reduce performance; vertebra is more robust since such perturbed views occur naturally in the dataset

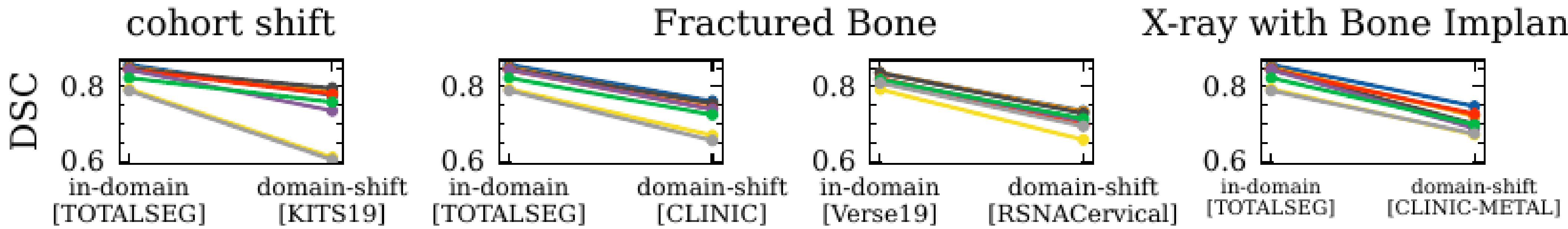


Performance gaps for clinical deployment

discrepancy in performance on clinically relevant minority subgroups

Reduced performance for Image & Population Domain shifts hinder clinical translation

Misaligned X-ray views can slightly reduce performance; vertebra is more robust since such perturbed views occur naturally in the dataset

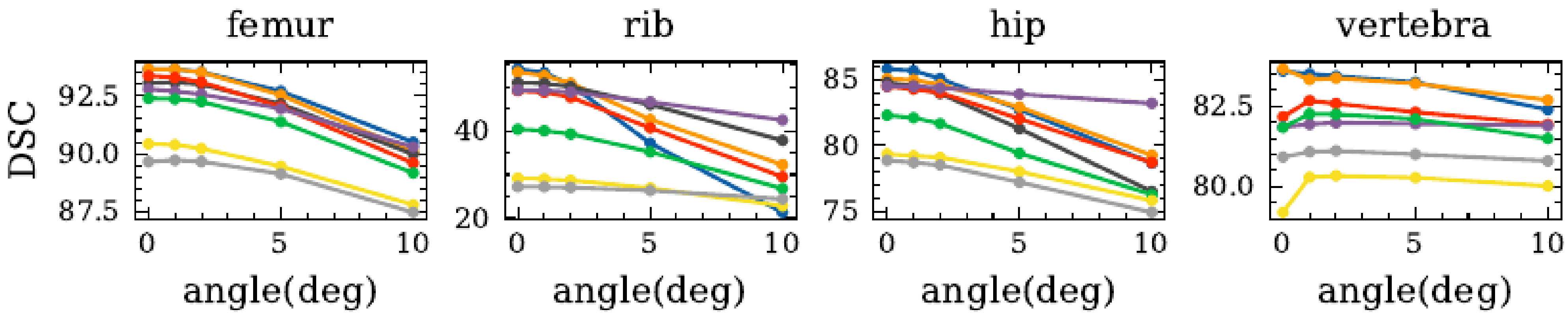


Performance gaps for clinical deployment

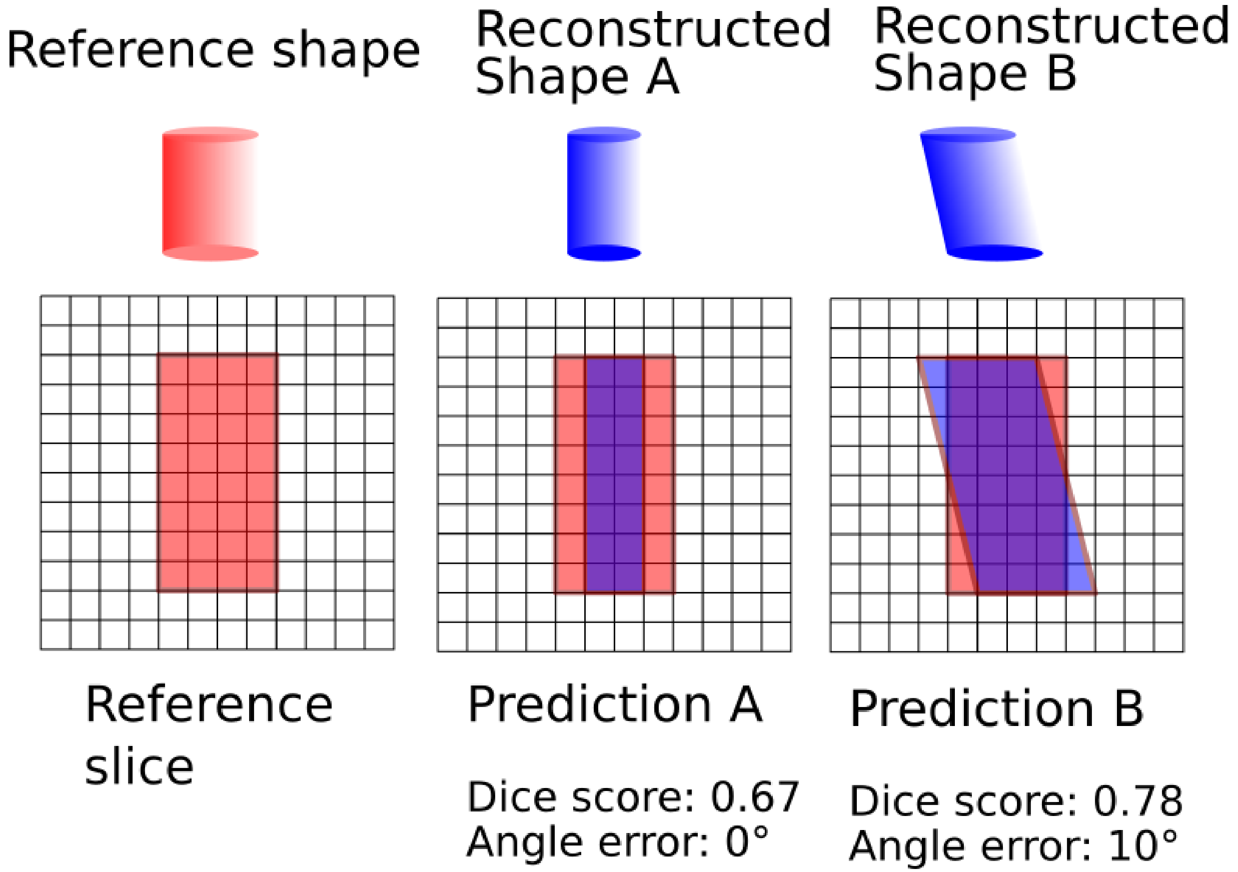
discrepancy in performance on clinically relevant minority subgroups

Reduced performance for Image & Population Domain shifts hinder clinical translation

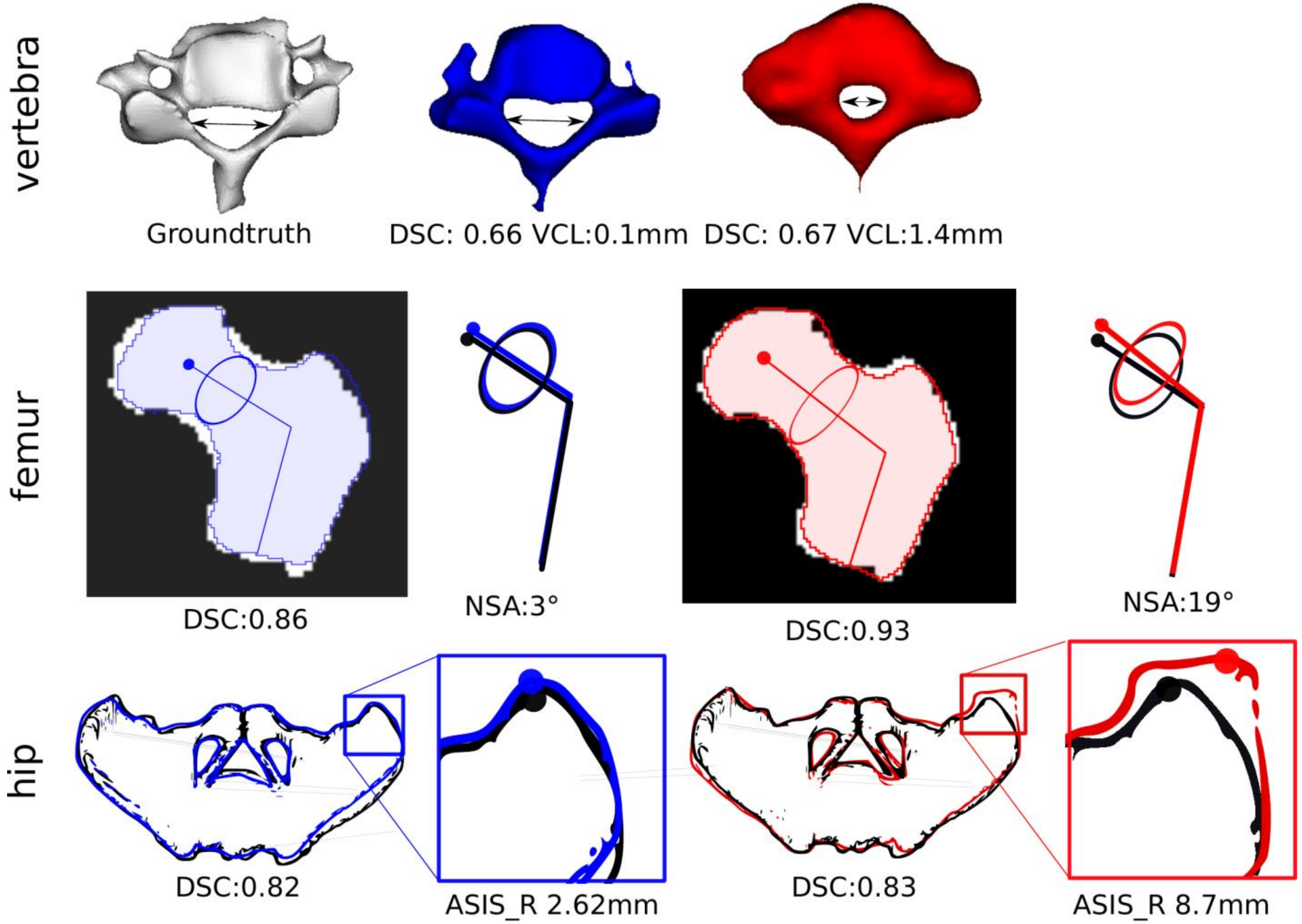
Misaligned X-ray views can slightly reduce performance; vertebra is more robust since such perturbed views occur naturally in the dataset



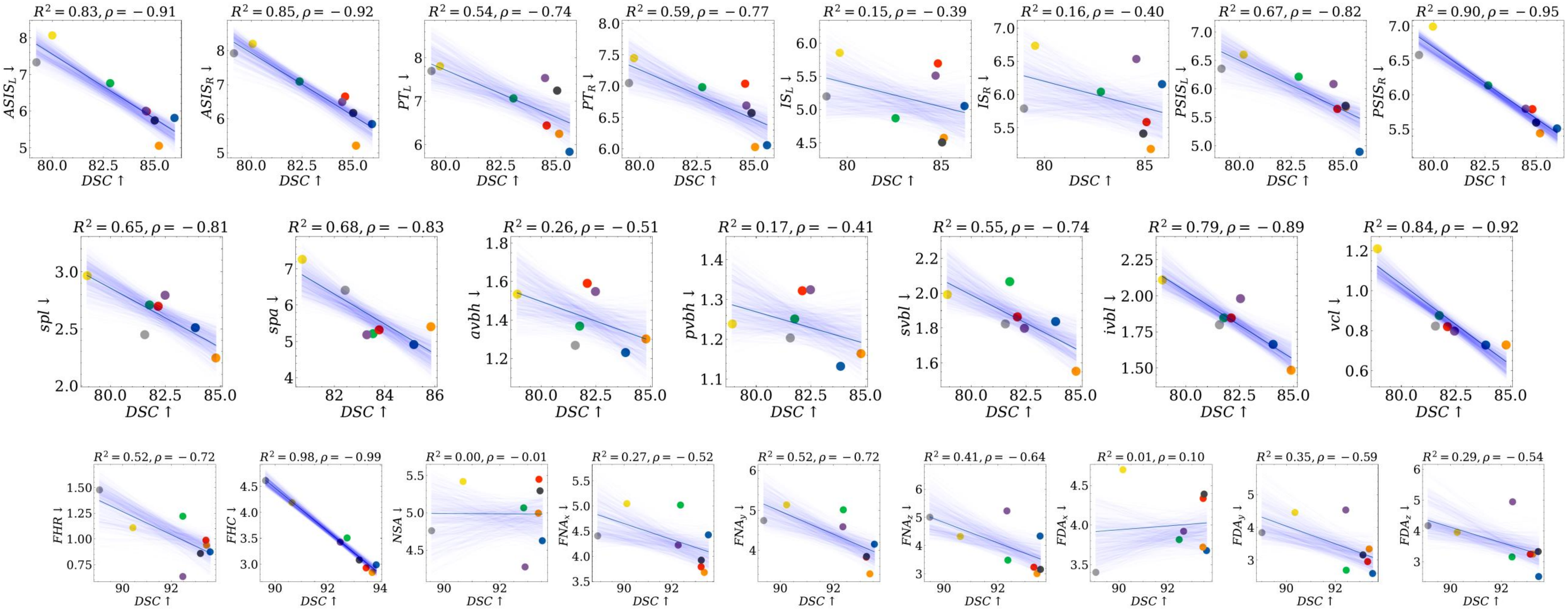
Dice Score alone is insufficient for clinical evaluation



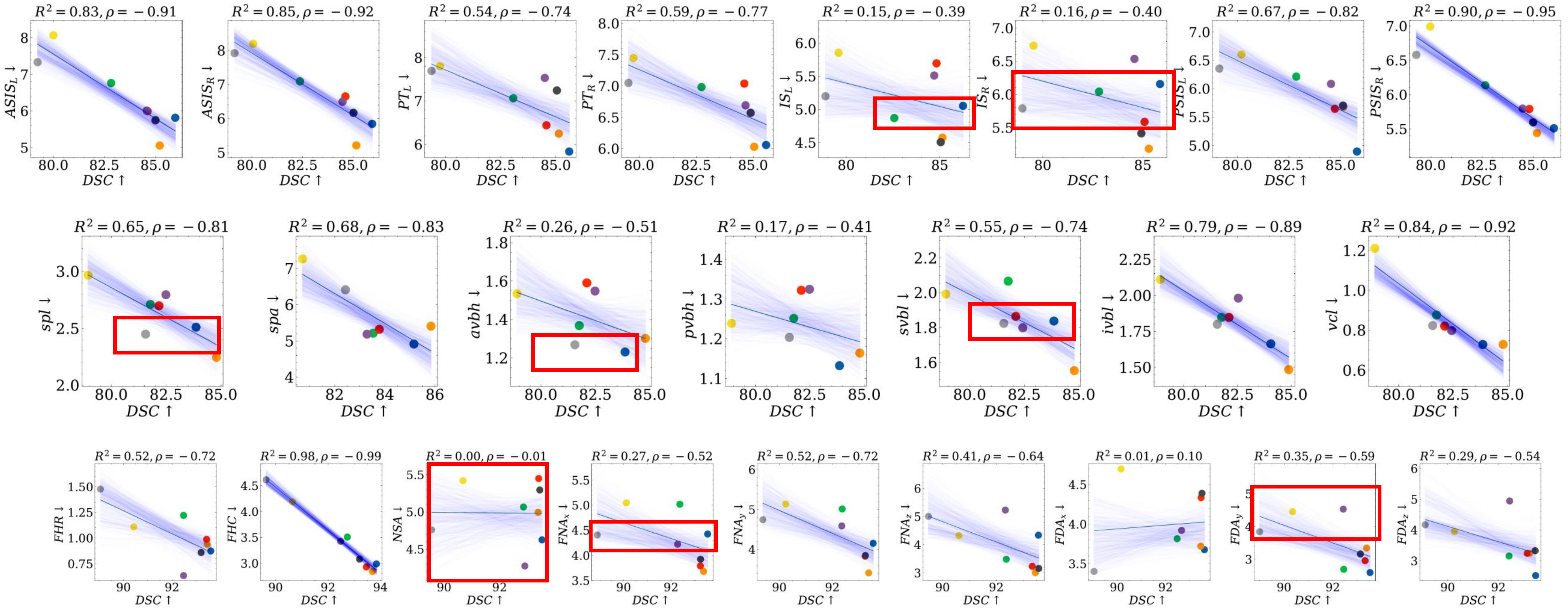
Dice Score alone is insufficient for clinical evaluation



Dice Score alone is insufficient for clinical evaluation



Dice Score alone is insufficient for clinical evaluation



Conclusion

First comprehensive benchmark and Evaluation of deep learning-based methods

Reproducible and accessible benchmark toolkit for the community to build upon

Overlap and Surface-based Metrics are not sufficient for clinical evaluation

Realistic Clinical Benchmarking Tasks show gaps in clinical adoption due to reduced performance on minority subgroups and domain shifts

Benchmark Toolkit: https://github.com/naamiinepal/xrayto3D-benchmark

LAT AP

Xray$_{to}$3DShape Benchmark