Beyond Landmarks: Anchor-Guided Orientation Regression for Robust Spinopelvic Measurement

Yunseob Hwang[1], Youngsuk Lee[1], DaeSeok Song[1], Dong-Ho Kang[2], and Du Hyun Ro[1,3,4*]

[1] CONNECTEVE Co., Ltd., Seoul, Korea
[2] Department of Orthopedic Surgery, Samsung Medical Center, Seoul, Korea
[3] College of Medicine, Seoul National University, Seoul, Korea
[4] Department of Orthopedic Surgery, Seoul National University Hospital, Seoul, Korea
duhyunro@gmail.com

Abstract. Spinopelvic measurement (SPM) plays a central role in spinal deformity assessment and surgical planning. Recent AI approaches predominantly adopt multi-stage pipelines that first localize vertebral regions then detect landmarks within isolated patches. While effective and stable, such decomposition restricts models to local subtasks rather than holistic reasoning. As a single-model alternative, direct landmark detection from full radiographs encounters two fundamental challenges: independent landmark prediction suffers from anatomical ambiguity leading to vertebral swapping, and point-wise localization objectives are misaligned with angular measurement goals. We reformulate SPM by replacing point-wise landmark detection with anchor-guided orientation regression that directly models task-relevant geometric primitives. Using frozen DINOv3 vision foundation model encoders with parameter-efficient fine-tuning, we systematically compare three representation methods under controlled experimental conditions—maintaining consistent architectures, decoders, and training protocols to isolate the effect of geometric output formulation. Evaluation on 309 external validation images demonstrates that anchor-guided orientation regression improves measurement consistency (ICC from 0.932 to 0.953) and accuracy (MAE from 2.32° to 2.18°). Critically, the representation maintains ICC above 0.9 even at 256-pixel input where baseline methods degrade significantly. These improvements hold across both ViT and ConvNeXt encoders. This study demonstrates that task-aligned geometric representation—not architectural complexity—is the key determinant of robust, resolution-tolerant automated spinopelvic measurement.

Keywords: Sagittal Balance · Spinopelvic Measurement · Landmark Detection · Orientation Prediction · Vision Foundation Model.

1 Introduction

Spinopelvic measurement (SPM) quantifies spinal and pelvic alignment from lateral lumbar radiographs—including lumbar lordosis (LL), L4-S1 lordosis, sacral

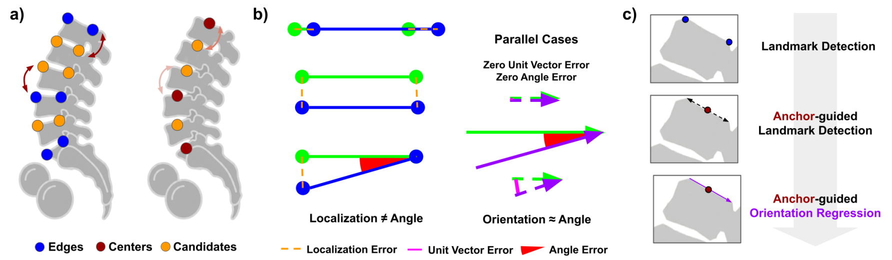

Fig. 1. Reformulating landmark detection for robust spinopelvic measurement. (a) Landmark prediction is reformulated in an anchor-centered reference frame to resolve anatomical ambiguity and suppress vertebral swapping. **(b)** As point-wise localization is often misaligned with clinical angular accuracy, orientation regression directly models the task-relevant geometric primitive, aligning optimization with the final measurement objective. **(c)** This progression reflects a shift from point-based landmarks to anchor-guided orientation regression (AGOR) for task-aligned geometric supervision.

slope (SS), pelvic incidence (PI), and pelvic tilt (PT)—and plays a critical role in orthopedic diagnosis and surgical planning [1, 2, 13, 19]. However, manual assessment remains subject to substantial inter- and intra-observer variability [3, 11, 12], limiting reproducibility and motivating automated solutions.

Recent AI approaches for SPM predominantly adopt top-down pipelines that decompose the task into vertebra localization, region cropping, and landmark detection within isolated regions [5, 7, 9, 18]. While effective in whole-spine scenarios—providing simplified context through cropping, sufficient resolution for localization, and improved overall pipeline stability—such decomposition introduces error propagation across stages and restricts models to solving local subtasks rather than reasoning over global anatomical relationships.

Direct landmark detection from the radiograph offers a single-stage alternative, yet it introduces structural ambiguity in vertebra-specific landmark detection: independently predicted vertebral edge points may be misassigned across adjacent levels, resulting in vertebral swapping (Fig. 1a). More fundamentally, conventional landmark regression trains models to improve localization accuracy, whereas SPM depends on vertebral orientation for angular computation, resulting in a misalignment between the optimization objective and the geometric primitive defining the task (Fig. 1b). While anchor-based representations have been explored for Cobb angle measurement in AP whole-spine radiographs—predicting vertebral centers with 4-corner offsets [23] or tilt angles [24]—their application to lateral lumbar SPM remains unexplored. Moreover, recent advances in vision foundation models (VFMs) such as DINOv3 [20] pretrained on 1.689 billion natural images have demonstrated remarkable performance across diverse vision tasks, including medical image analysis [16, 14]. However, their application to anatomical landmark detection and angle measurement in SPM remains unexplored, requiring systematic experimental validation of VFM-based architectures for this clinical task.

To address these gaps, we investigate representation reformulation as the central design variable for SPM. In this study, we use lateral lumbar radiographs as a controlled setting to systematically compare Landmark Detection, Anchor-Guided Landmark Detection (AGLD), and Anchor-Guided Orientation Regression (AGOR). All methods are implemented under standardized model architectures comprising a DINOv3 encoder, a lightweight heatmap decoder, and task-specific geometric prediction heads, with identical training protocols across representations. This progression from point-based landmarks to anchor-centered orientation primitives is illustrated in Fig. 1c. By modifying only the geometric output formulation while keeping over 99% of the model architecture unchanged, we isolate the impact of representation choice on clinical angle measurement.

Our results demonstrate that task-aligned orientation regression improves angular accuracy and resolution robustness while simplifying pipeline design, highlighting representation alignment—not architectural complexity—as a key determinant of robust spinopelvic measurement.

2 Method

2.1 Dataset and Annotation

We curated 1,209 lateral lumbar spine X-rays from two institutions: Seoul National University Hospital (SNUH) (n = 900; 720 training, 180 validation) and Samsung Medical Center (SMC) (n = 309 external validation set). All images capture the anatomical region from lumbar vertebrae 1 (L1) through sacrum (S1) with bilateral femoral heads visible. A board-certified orthopedic surgeon with over 10 years of experience annotated 8 anatomical landmarks on each image: 6 vertebral superior endplate points (anterior and posterior for L1, L4, and S1) and 2 femoral head centers. Femoral head centers were derived through automated ellipse fitting followed by manual correction when necessary to ensure anatomical accuracy. Training and validation splits were stratified to maintain consistent parameter distributions across subsets.

2.2 Methodologies

We compare three representation methodologies (Fig. 2) that differ only in their geometric output formulation while sharing identical DINOv3 encoder architectures: Landmark Detection (8 independent landmarks), AGLD (5 centers + 12 pixel offsets encoding anterior and posterior endpoints for 3 vertebrae), and AGOR (5 centers + 6 unit orientation vectors encoding posterior directions for 3 vertebrae).

Landmark Detection (Baseline) The baseline method predicts 8 independent heatmaps (one per landmark) using Gaussian kernels centered at ground truth coordinates. Peak locations are extracted via argmax operation, and vertebral orientation vectors are computed post-hoc by connecting anterior-posterior endplate point pairs. This independent prediction strategy provides no explicit constraints to prevent vertebral misassignment, where predicted landmarks may be incorrectly associated with adjacent vertebral levels (e.g., L1 anterior landmark assigned to thoracic vertebra T12).

Anchor-Guided Landmark Detection (AGLD) AGLD decomposes the task into anatomical grouping by predicting 5 center heatmaps (L1, L4, S1 vertebral body centers, and bilateral femoral head centers) plus 12 auxiliary offset channels. Each offset channel encodes pixel displacements $(\Delta x, \Delta y)$ from vertebral centers to their corresponding anterior and posterior endplate points. Final landmark coordinates are reconstructed via:

$$\hat{P} = \hat{C}_i + \hat{\Delta}(\hat{C}_i) \tag{1}$$

where $\hat{C}_i$ is the detected center and $\hat{\Delta}(\hat{C}_i)$ is the predicted offset sampled at that center location. This anchor-guided formulation introduces structural constraints that reduce vertebral misassignment by explicitly grouping landmarks according to anatomical hierarchy.

Anchor-Guided Orientation Regression (AGOR) AGOR reformulates the representation to directly predict task-relevant geometric primitives. Instead of regressing pixel coordinates, the model predicts 5 center heatmaps plus 6 orientation channels encoding unit direction vectors from vertebral centers toward posterior endplates (a single direction suffices to define vertebral orientation). Ground truth orientation vectors are computed as:

$$\mathbf{u}_{\text{post}} = \frac{P_{\text{post}} - C_i}{\|P_{\text{post}} - C_i\|} \tag{2}$$

where P_{post} is the posterior endplate point and C_i is the vertebral center. Each orientation is encoded as a 2-channel field (u_x, u_y) representing the unit vector components. During inference, predicted orientation vectors $\hat{\mathbf{u}}_{\text{post}}$ are sampled at detected center locations $\hat{C}_i$ and used directly for angle calculation without intermediate coordinate reconstruction. This representation alignment eliminates the objective mismatch between pixel-level localization and angular measurement, enabling the model to learn vertebral orientations—the actual geometric primitive used in clinical parameter computation.

2.3 Model Architecture

The model employs a frozen DINOv3 encoder [20], a decoder network, and task-specific prediction heads (Fig. 2). We evaluate two encoder architectures: ViT-B/16 [4] (85.7M parameters) paired with simple deconvolution layers [22] (×4

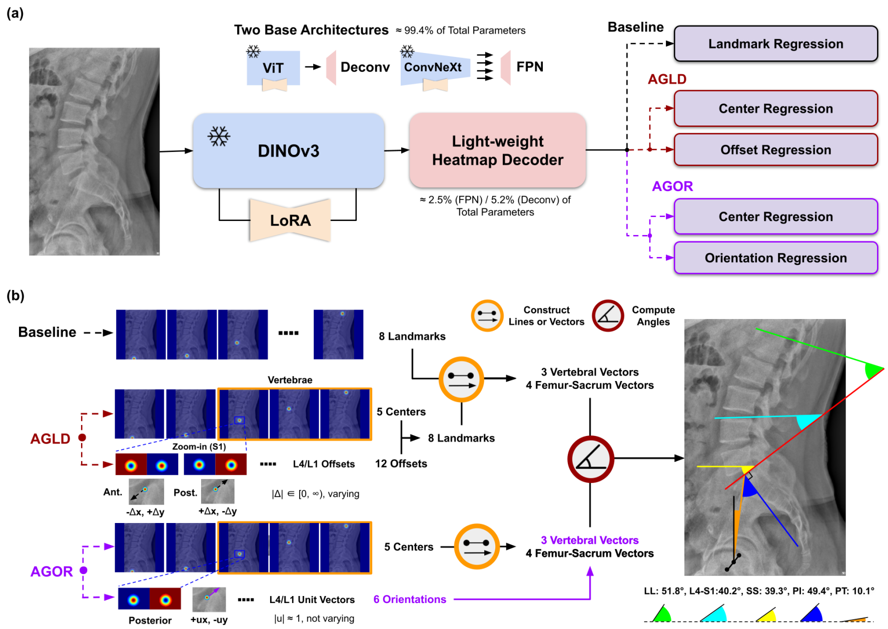

Fig. 2. Model Architecture Overview and Angle Measurement Pipeline. (a) Architecture overview: All methods share DINOv3 encoder + decoder (>99% of parameters), differing only in task-specific prediction heads (<1%). **(b)** Prediction outputs: Baseline predicts 8 landmarks; AGLD predicts 5 centers + 12 pixel offsets; AGOR predicts 5 centers + 6 unit vectors for direct angle calculation.

upsampling from 1/16 resolution last layer feature), and ConvNeXt-Base [15] (87.6M parameters) with Feature Pyramid Network (FPN, multi-scale features at 1/32, 1/16, 1/8, 1/4 resolutions). During training, we apply Low-Rank Adaptation (LoRA) with rank 8 and alpha 16 for parameter-efficient fine-tuning. Both decoders produce 1/4 resolution heatmaps with 256-channel feature maps. Prediction heads vary by method, using either a single landmark head (Landmark Detection) or a dual-branch architecture with landmark and auxiliary heads (AGLD and AGOR).

2.4 Learning Objectives

Landmark Detection uses MSE loss for 8 heatmaps:

$$\mathcal{L}_{\text{heatmap}} = \frac{1}{N}\sum_{i=1}^{8} \|\mathbf{H}_i - \hat{\mathbf{H}}_i\|^2 \tag{3}$$

AGLD and AGOR use multi-task loss combining MSE for center heatmaps and Smooth L1 for offset/orientation fields:

$$\mathcal{L} = \lambda_{\text{center}}\mathcal{L}_{\text{MSE}}(\mathbf{H}) + \lambda_{\text{reg}}\mathcal{L}_{\text{SmoothL1}}(\mathbf{O}) \tag{4}$$

with $\lambda_{\text{center}} = 1.0$, $\lambda_{\text{reg}} = 0.5$ (AGLD) and 60.0 (AGOR, scaled to match offset magnitude).

2.5 Angle Calculation

All five spinopelvic parameters are derived from vertebral orientations and femoral head midpoint: LL (L1-L4 endplate angle), L4-S1 (L4-S1 lordosis), SS (S1 inclination to horizontal), PI (perpendicular to S1 endplate vs. line to femoral midpoint), PT (vertical vs. line to S1 midpoint). Landmark Detection and AGLD reconstruct angles from coordinates; AGOR directly predicts vertebral orientations, bypassing point-to-line reconstruction and mitigating error propagation in angle measurement.

3 Experiments and Results

3.1 Implementation Details

Models were implemented in PyTorch 2.0 with mixed-precision training on a single NVIDIA L40S 48GB GPU. We applied LoRA [8] (rank 8, alpha 16) to attention and feed-forward layers (ViT-B/16) or pointwise convolution layers (ConvNeXt-Base), resulting in 3.7M (4.1%) and 5.0M (5.4%) trainable parameters respectively. Ground truth heatmaps used resolution-dependent Gaussian kernels ($\sigma = 6/4.5/3/1.5$ for 1024/768/512/256px). Images were resized to target resolution preserving aspect ratio with zero-padding. Training used AdamW optimizer ($\eta = 10^{-4}$, 100 epochs, cosine annealing) with augmentation including horizontal flipping, rotation ($\pm 5°$), brightness/contrast adjustment, and noise injection. Model selection was based on best validation loss.

3.2 Evaluation Metrics

We evaluated landmark localization using Mean Radial Error (MRE, in millimeters, converted from pixels using 0.14 mm/px) and Success Detection Rate at 2mm threshold (SDR@2mm, percentage of landmarks within 2mm). Clinical angle accuracy was measured by Mean Absolute Error (MAE, in degrees) and ICC (values > 0.9 indicate excellent agreement [10]). Avg MAE and Avg ICC denote the mean across all five spinopelvic angles.

3.3 Quantitative Results

Table 1 compares three representation methods using two DINOv3 encoders on the external validation set ($n = 309$). AGOR consistently outperformed baselines, achieving ICC 0.9532 and MAE 2.18° with ConvNeXt-Base. The anchor-guided orientation formulation aligns supervision with the clinical task—learning vertebral endplate orientation rather than pixel-level landmark positions. This representation alignment improved average ICC from 0.9320 to 0.9532 (ConvNeXt-Base) and 0.9320 to 0.9444 (ViT-B/16).

Table 1. Comparison of Representation Methods and DINOv3 Encoders

Method	DINOv3 Encoder	**Landmark Localization**		**Angle Measurement**	
		MRE (mm)	SDR@2mm	Avg MAE (°)	Avg ICC
Landmark Detection	ViT-B/16	2.62	53.03	2.68	0.9320
Landmark Detection	ConvNeXt-B	2.46	61.77	2.32	0.9308
AGLD	ViT-B/16	3.12	39.52	2.61	0.9438
AGLD	ConvNeXt-B	2.48	56.11	2.33	0.9526
AGOR	ViT-B/16	2.74	45.50	2.64	0.9444
AGOR	**ConvNeXt-B**	**2.05**	**67.12**	**2.18**	**0.9532**

Comparison with Previous Works. Table 2 compares AGOR against prior automated spinopelvic measurement systems. Our single end-to-end model achieves consistently higher ICC values across all five parameters on the largest external validation set ($n = 309$), while previous approaches require 2–3 specialized models with smaller test sets (40–118 samples).

Table 2. ICC Comparison with Previous Automated Spinopelvic Measurement Systems

Study	**# Models**	**# Test Samples**	**LL**	**L4-S1**	**SS**	**PI**	**PT**
Orosz et al. [18]	2	100 (Pre-op)	0.92	–	0.88	0.88	0.85
Song et al. [21]	1	90	0.78	–	0.62	0.78	0.85
Löchel et al. [17]	3	118 (Pre-op)	0.81	–	0.71	0.72	0.85
Harake et al. [6]	3	40	0.94	–	0.93	0.91	0.98
Hoppe et al. [7]	2	65	0.95	–	0.93	–	–
Ours (AGOR)	**1**	**309**	**0.97**	**0.95**	**0.94**	**0.93**	**0.98**

Bland–Altman Analysis. Figure 3 presents Bland–Altman plots confirming strong AI–expert agreement for the best-performing model (AGOR, ConvNeXt-Base), with mean bias below 0.7° and limits of agreement within ±7° across all five spinopelvic parameters.

Ablation Study - Resolution Robustness. Figure 4 shows performance across four resolutions. AGOR maintains ICC>0.90 even at 256px while baselines degrade below this threshold.

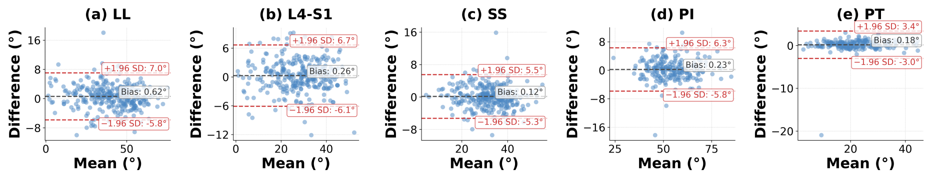

Fig. 3. Bland–Altman Analysis. AI vs. expert agreement on the external validation set.

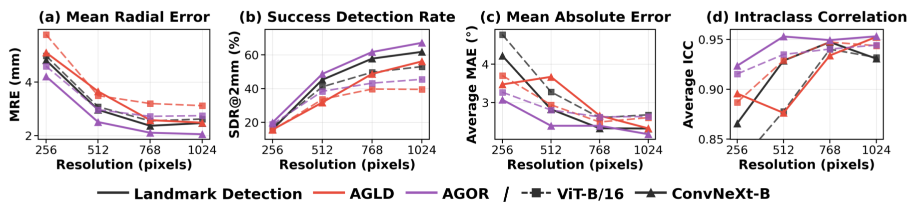

Fig. 4. Resolution Robustness. Performance at 256–1024 pixels for three methods with two backbones (ViT-B/16 dashed, ConvNeXt-Base solid). **(a)** MRE, **(b)** SDR@2mm, **(c)** MAE, **(d)** ICC. AGOR maintains superior performance across all resolutions.

3.4 Qualitative Results

Figure 5 presents two representative failure cases. Landmark Detection exhibits vertebral swapping (L1 anterior misassigned to thoracic vertebra above) and both baselines mispredict L1 landmarks at L2 level, while AGOR maintains accuracy in both cases through direct orientation prediction, confirming robustness to anatomical ambiguity.

4 Conclusion

This work demonstrates that representation reformulation, rather than architectural complexity, is the key determinant of robust angle measurement in spinopelvic assessment. Comparing landmark detection, AGLD, and the proposed AGOR, we show that while anchor guidance addresses anatomical ambiguity, offset-based coordinate regression retains the mathematical mismatch between training objectives and clinical goals—which AGOR resolves by directly predicting task-relevant orientation primitives. Critically, by modifying only the task-specific prediction heads ($< 1\%$ of total parameters), AGOR improves average ICC from 0.932 to 0.953 and reduces MAE from 2.32° to 2.18° compared to the landmark detection baseline, while maintaining ICC>0.9 even at 256px where baselines degrade. These findings establish direct orientation regression as the key to accurate and robust angle measurement.

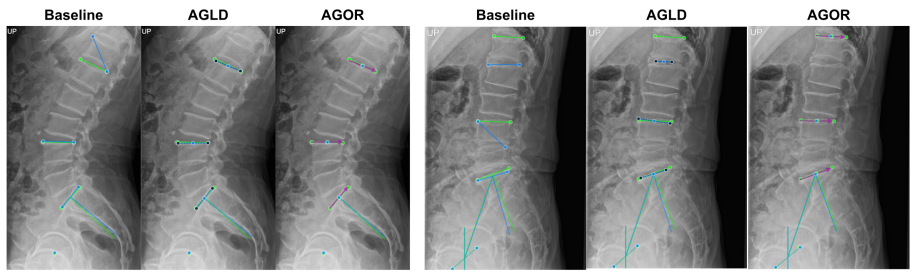

Fig. 5. Qualitative Comparison. Two representative cases comparing three methods. AGOR achieves accurate measurements while Landmark Detection (Baseline) shows vertebral swapping and AGLD exhibits localization errors propagating to angle calculation.

Acknowledgments. This study was approved by the Institutional Review Board of Seoul National University Hospital (IRB No. 2009-181-1161) and Samsung Medical Center (IRB No. 2024-12-119), with informed consent waived due to the retrospective nature of the study. The authors thank the clinical and annotation teams at both institutions for their contributions to data curation and quality assurance.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ames, C.P., Smith, J.S., Scheer, J.K., Bess, S., Bederman, S.S., Deviren, V., et al.: Impact of spinopelvic alignment on decision making in deformity surgery in adults: a review. Journal of Neurosurgery: Spine **16**(6), 547–564 (2012)
2. Diebo, B.G., Varghese, J.J., Lafage, R., Schwab, F.J., Lafage, V.: Sagittal alignment of the spine: What do you need to know? Clinical Neurology and Neurosurgery **139**, 295–301 (2015)
3. Dimar, J.R., Carreon, L.Y., Labelle, H., Djurasovic, M., Weidenbaum, M., Brown, C., Roussouly, P.: Intra- and inter-observer reliability of determining radiographic sagittal parameters of the spine and pelvis using a manual and a computer-assisted methods. European Spine Journal **17**(10), 1373–1379 (2008)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
5. Grover, P., Siebenwirth, J., Caspari, C., Drange, S., Dreischarf, M., Le Huec, J., Putzier, M., Franke, J.: Can artificial intelligence support or even replace physicians in measuring sagittal balance? A validation study on preoperative and postoperative full spine images of 170 patients. European Spine Journal **31**, 1943–1951 (2022)
6. Harake, E.S., Linzey, J.R., Jiang, C., Joshi, R.S., Zaki, M.M., Jones, J.C., Khalsa, S.S., Lee, J.H., Wilseck, Z., Joseph, J.R., Hollon, T.C., Park, P.: Development and validation of an artificial intelligence model to accurately predict spinopelvic parameters. Journal of Neurosurgery: Spine, 1–9 (2024)

7. Hoppe, B.F., Rueckel, J., Rudolph, J., Fink, N., Weidert, S., Hohlbein, W., et al.: Automated spinopelvic measurements on radiographs with artificial intelligence: a multi-reader study. La Radiologia Medica **130**(3), 359–367 (2025)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
9. Kang, D., Jeong, Y., Kim, S.T., Kim, Y., Chang, B., Kim, H., Chang, S.Y., Ro, D.: Automated measurement of pelvic parameters using convolutional neural network in complex spinal deformities: Overcoming challenges in coronal deformity cases. The Spine Journal (2025)
10. Koo, T.K., Li, M.Y.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine **15**(2), 155–163 (2016)
11. Kyrölä, K.K., Salme, J., Tuija, J., Tero, I., Eero, K., Arja, H.: Intra- and interrater reliability of sagittal spinopelvic parameters on full-spine radiographs in adults with symptomatic spinal disorders. Neurospine **15**(2), 175 (2018)
12. Lafage, R., Ferrero, E., Henry, J.K., Challier, V., Diebo, B., Liabaud, B., Lafage, V., Schwab, F.: Validation of a new computer-assisted tool to measure spino-pelvic parameters. The Spine Journal **15**(12), 2493–2502 (2015)
13. Le Huec, J.C., Faundez, A., Dominguez, D., Hoffmeyer, P., Aunoble, S.: Evidence showing the relationship between sagittal balance and clinical outcomes in surgical treatment of degenerative spinal diseases: a literature review. International Orthopaedics **39**(1), 87–95 (2015)
14. Li, Y., Wu, Y., Lai, Y., Hu, M., Yang, X.: MedDINOv3: How to adapt vision foundation models for medical image segmentation? arXiv preprint arXiv:2509.02379 (2025)
15. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
16. Liu, C., Chen, Y., Shi, H., Lu, J., Jian, B., Pan, J., Cai, L., Wang, J., Yu, J., Gao, Z., Zhang, X.: Does DINOv3 set a new medical vision standard? Benchmarking 2D and 3D classification, segmentation, and registration. arXiv preprint arXiv:2509.06467 (2025)
17. Löchel, J., Eckstein, F., Babawale, L., Palmowski, Y., Stüber, A.T., Huppertz, A., et al.: Deep learning algorithm for fully automated measurement of sagittal balance in adult spinal deformity. European Spine Journal **33**(11), 4119–4124 (2024)
18. Orosz, L.D., Bhatt, F.R., Jazini, E., Dreischarf, M., Grover, P., Grigorian, J.N., Roy, R.T., Schuler, T.C., Good, C.R., Haines, C.M.: Novel artificial intelligence algorithm: an accurate and independent measure of spinopelvic parameters. Journal of Neurosurgery: Spine, 1–9 (2022)
19. Schwab, F., Patel, A., Ungar, B., Farcy, J.P., Lafage, V.: Adult spinal deformity—postoperative standing imbalance: how much can you tolerate? An overview of key parameters in assessing alignment and planning corrective surgery. Spine **35**(25), 2224–2231 (2010)
20. Simeoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jegou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025)

21. Song, S.Y., Seo, M.S., Kim, C.W., Kim, Y.H., Yoo, B.C., Choi, H.J., et al.: AI-driven segmentation and automated analysis of the whole sagittal spine from X-ray images for spinopelvic parameter evaluation. Bioengineering **10**(10), 1229 (2023)
22. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 38571–38584 (2022)
23. Yi, J., Wu, P., Huang, Q., Qu, H., Metaxas, D.N.: Vertebra-focused landmark detection for scoliosis assessment. In: IEEE International Symposium on Biomedical Imaging (ISBI). pp. 736–740 (2020)
24. Zou, L., Guo, L., Zhang, R., Ni, L., Chen, Z., He, X., Wang, J.: VLTENet: A deep-learning-based vertebra localization and tilt estimation network for automatic Cobb angle estimation. IEEE Journal of Biomedical and Health Informatics **27**, 3002–3013 (2023)