This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

A Dataset

The Genodisc grading distribution is summarized in Table A1. Pfirrmann, the only five-grade task, comprises Grade I: 3911 (32.3%), Grade II: 1763 (14.6%), Grade III: 2824 (23.4%), Grade IV: 2424 (20.0%), and Grade V: 1169 (9.7%).

Table A1: Pathology grade frequency per grading task for the complete Genodisc dataset (all splits).

Task	Normal (%)	Mild (%)	Moderate (%)	Severe (%)
Narrowing	7246 (60.0)	1387 (11.5)	2198 (18.2)	1242 (10.3)
Canal stenosis	11364 (94.1)	292 (2.4)	186 (1.5)	236 (2.0)
Upper Endplate defect	11045 (91.4)	950 (7.9)	61 (0.5)	26 (0.2)
Lower Endplate defect	11023 (91.3)	921 (7.6)	88 (0.7)	47 (0.4)
Foraminal stenosis Left	10867 (89.9)	676 (5.6)	275 (2.3)	265 (2.2)
Foraminal stenosis Right	10841 (89.7)	753 (6.2)	244 (2.0)	243 (2.0)
Herniation	10758 (89.0)	639 (5.3)	424 (3.5)	262 (2.2)
Bulging (Ant.)	8953 (74.1)	2050 (17.0)	765 (6.3)	315 (2.6)
Bulging (Post.)	8451 (69.9)	2607 (21.6)	828 (6.9)	198 (1.6)
Spondylolisthesis	11609 (96.2)	434 (3.6)	30 (0.2)	7 (0.1)

B Auxiliary Segmentation Loss and Continuous Ranking

Auxiliary loss. Table B1 summarizes the effect of keeping the segmentation loss active (λ_{seg}=0.5) as an auxiliary term during fine-tuning. Gains are small but most pronounced in regimes with low percentages of grading supervision.
Continuous ranking. The embedding z also yields a continuous severity score $s_t \in [0, S]$ read from the task logits, supervised by a severity-weighted hinge [1, 13] over differently-graded pairs $\mathcal{P}_t$,

$$\mathcal{L}^t_{\text{hinge}} = \frac{1}{|\mathcal{P}_t|} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{P}_t} w^t_{ij} \max\big(0,\, m^t_{ij} - r^t_{ij}(s_{ti} - s_{tj})\big), \tag{1}$$

with $r^t_{ij} \in \{-1, 0, 1\}$, weight $w^t_{ij} \propto |y_{ti} - y_{tj}|/(K_t{-}1)$ and margin $m^t_{ij} = m_0 + m_1|y_{ti} - y_{tj}|/(K_t{-}1)$ (m_0=2.0, m_1=1.5), plus a similarity term $\mathcal{L}^t_{\text{sim}} = \frac{1}{N_t}\sum_i (s_{ti} - y_{ti}S/(K_t{-}1))^2$ that ties each score to $y_{ti}S/(K_t{-}1)$, giving within-grade compactness and inter-grade calibration. We refer to [13] for the full derivation of the continuous score. With $\mathcal{L}_{\text{rank}} = \sum_t (\mathcal{L}^t_{\text{hinge}} + \lambda \mathcal{L}^t_{\text{sim}})$, λ=0.25, segmentation pre-training leads on cumulative ROC-AUC, QWK (Table B2).

Table B1: Mean one-vs-rest ROC-AUC by auxiliary segmentation-loss weight λ_{seg}.

	Scratch			Seg-pretrained		
λ_{seg}	20%	50%	100%	20%	50%	100%
0	0.875	0.916	0.917	0.911	0.924	0.928
0.5	0.895	0.919	0.924	0.924	0.935	0.942

Table B2: Macro cumulative ROC-AUC and QWK from the continuous severity score.

	Scratch			Seg-pretrained		
Metric	20%	50%	100%	20%	50%	100%
AUC	0.941	0.953	0.959	**0.956**	**0.960**	**0.961**
QWK	0.629	0.671	0.691	**0.688**	**0.708**	**0.717**

C Morphological Parameter Extraction

From the predicted masks we extract interpretable geometric descriptors (mm) by fitting each structure's principal axes: disc heights perpendicular to the major axis, canal AP width along the minor axis, and maximum disc bulge as the largest radial distance the boundary extends past a fitted ellipse. Associations with grade use Spearman's ρ (Fig. C1). As descriptors are non-normal (Shapiro–Wilk $p < 0.05$), grade separation is assessed non-parametrically: a Kruskal–Wallis test per parameter, then adjacent-grade Mann–Whitney U tests, both FDR-corrected (Benjamini–Hochberg).

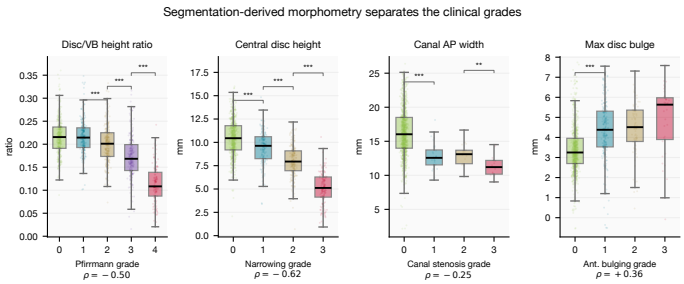

Fig. C1: Segmentation-derived geometrical parameters by ground-truth grade, extracted on the test set. Panels show the four descriptors with the strongest grading associations. Axes report Spearman ρ. Significant adjacent-grade differences (Mann–Whitney U) corrected (Benjamini–Hochberg FDR) are marked ($^*q < 0.05$, $^{**}q < 0.01$, $^{***}q < 0.001$).