Endplate3D-QCT: A High-Resolution Dataset and Benchmark for Automated 3D Segmentation of Lumbar Vertebral Endplates in QCT

Zixun Yin[1], Da Zou[2], Yi Zhao[2], Chenbin Zhang[3], Weishi Li[2], Minghui Wu[1,6], Kun Yan[3,4*], and Ping Wang[1,4,5*]

[1] School of Software and Microelectronics, Peking University, Beijing, China
[2] Department of Orthopaedics, Peking University Third Hospital, Beijing, China
[3] School of Computer Science, Peking University, Beijing, China
[4] National Engineering Research Center for Software Engineering, Peking University, Beijing, China
[5] Key Laboratory of High Confidence Software Technologies (PKU), Ministry of Education, China
[6] Shanghai Artificial Intelligence Laboratory and Mininglamp Technology, China

Abstract. Accurate segmentation of lumbar vertebral endplates is essential for assessing bone density and biomechanical properties in spinal disorders. While quantitative computed tomography (QCT) provides detailed bone density measurements, existing segmentation approaches primarily focus on vertebral bodies and intervertebral discs, often neglecting the precise delineation of endplates. Current deep learning methods perform well in healthy spines but struggle with pathological cases due to the thin and morphologically complex nature of endplates, particularly in the presence of osteophytes and degenerative changes. To address these challenges, we introduce the first publicly available dataset, Endplate3D-QCT, which contains pixel-level annotations of lumbar endplates in clinical QCT scans. Our dataset includes high-precision 3D segmentation masks targeting cortical endplates and subchondral bone, along with an automated evaluation framework for model assessment. We benchmark multiple deep learning models, including EfficientUNet, UNet, VNet, UNETR and SwinUNETR, using nnUNet as the training framework. While these models achieve Dice scores around 0.9, they exhibit inconsistencies in endplate identification, leading to false positives and false negatives. These findings highlight the need for further advancements in endplate segmentation techniques. Our dataset and benchmarks provide a valuable foundation for improving spinal implant design, bone density mapping, and computational modeling of vertebral load distribution. The dataset and the evaluation code are available at https://github.com/yin876705249/Endplate3D-QCT.

Keywords: Lumbar Endplate Segmentation · Open-access Dataset · Evaluation Framework

* Corresponding authors: Kun Yan (kyan2018@pku.edu.cn) and Ping Wang (pwang@pku.edu.cn).

1 Introduction

Accurate segmentation of lumbar vertebral endplates is critical for quantitative assessment of bone density and biomechanical characterization in spinal disorders. While quantitative computed tomography (QCT) enables density measurements of trabecular and cortical bone, existing segmentation methods for spinal structures—typically focusing on intervertebral discs and vertebral bodies—fail to achieve the precision required for endplate-specific analysis. Current deep learning approaches often yield satisfactory results in healthy spines with minimal osteophytes but struggle to delineate the thin (<4 mm), anatomically ambiguous endplate regions in pathological cases. This limitation stems from two key factors: 1) the reliance on threshold-based 2D methods that inadequately capture 3D morphological variations of the endplate's cortical bone layer, and 2) the lack of high-resolution CT datasets with expert annotations specifically targeting the osseous endplate-cartilage interface.

The clinical imperative of endplate segmentation lies in its direct relevance to spinal surgery outcomes. During procedures like interbody fusion, where implants contact the endplate surface, local bone density measurements could guide personalized device selection and 3D-printed cage design—critical for preventing subsidence in osteoporotic patients. Furthermore, precise endplate characterization could advance research on asymmetric bone remodeling in scoliosis and degenerative disc disease. While MRI excels at visualizing soft tissues like discs and neural structures [3], CT remains the modality of choice for osseous endplate analysis due to superior spatial resolution and contrast for cortical bone.

Despite these applications, no publicly available dataset currently provides detailed 3D annotations of lumbar endplates. Existing spinal CT segmentation works [12, 10] predominantly focus on vertebral bodies and discs, often treating the endplate as a byproduct of post-processing rather than a distinct anatomical entity. This gap hinders the development of machine learning models capable of capturing the endplate's subtle radiological signatures, particularly in degenerative spines where pathological changes obscure traditional intensity-based segmentation boundaries.

To address these challenges, we present the first open-access dataset with pixel-level annotations of lumbar endplates in clinical QCT scans. Our contributions include: 1) High-precision 3D segmentation masks delineating cortical endplates and subchondral bone regions. 2) An automated evaluation framework for precise assessment of segmentation model performance. 3) Comprehensive benchmarking experiments comparing widely-used segmentation methods.

This resource aims to accelerate research in personalized spinal implant design, bone density mapping, and computational modeling of load distribution across the vertebral column. We conducted benchmark experiments based on nnUNet [6, 7], a state-of-the-art framework in the medical image segmentation domain. We evaluated the performance of widely used segmentation methods, including UNet [11], VNet [9], and SwinUNETR [4], etc. The results indicate that while these methods achieve high segmentation performance (with Dice scores around 0.9), they do not consistently ensure the accurate identification

of all endplates, leading to instances of false negatives and false positives. This suggests that endplate segmentation remains a challenging task that requires further exploration.

2 Dataset Construction

2.1 Data Acquisition

Our study utilized 119 CT volumes from the CTSpine1K dataset [2] containing complete lumbar regions, which were carefully selected based on image quality and anatomical completeness. After extracting volumetric regions of interest (ROIs) centered on the lumbar spine, we performed pixel-level 3D annotations of vertebral endplates following clinical guidelines, including superior/inferior endplates from L1 to L5 and the superior endplate of S1. This resulted in 1,309 annotated endplate surfaces (11 per case) with sub-millimeter precision, addressing the critical gap in publicly available endplate segmentation data. To our knowledge, this represents the first open-access dataset providing large-scale (over 1,000 annotated surfaces), high-quality 3D annotations of lumbar endplates, particularly valuable for developing AI-assisted surgical planning systems and biomechanical modeling in spinal disorders.

In addition to the re-annotated CTSpine-Refined dataset, we collected another 51 patients' preoperative lumbar CT images in this study. These patients were clinically diagnosed as lumbar degenerative diseases and underwent lumbar fusion surgery. The scanning parameters of CT were set at the values of 120 kVp. The patients group included 23 males and 28 females, and their average age was 66.2 ± 5.3 years, ranging from 51 to 76 years. Concretely, we constructed two datasets from these 51 clinical cases (10 with scoliosis) of lumbar degenerative diseases (LDD) reflecting different levels of segmentation difficulty:

1. *LDD-Mild* includes 25 cases with mild symptoms and regular endplates, representing a relatively low level of segmentation complexity.
2. *LDD-Severe* includes 26 cases with severe symptoms and irregular endplates, posing significant challenges for segmentation models. These datasets serve to test model robustness across a spectrum of anatomical variations.

In this study, the cortical bone of upper and lower lumbar endplates from L1 to L5 together with S1 upper endplates were segmented on lumbar CT images. The segmentation process was performed in the sagittal plane of CT images slice by slice to annotate the whole endplate, as shown in Fig. 1. Five spine surgeons participated in the CT image annotation. Four surgeons with over 5 years of clinical experience performed the initial segmentation. All received standardized training and strictly followed the annotation manual to minimize bias. Using cross-annotation and iterative feedback, the team held multiple discussions to resolve inconsistencies. Additionally, a spine surgeon with more than 10 years of experience reexamined all annotated images, further ensuring quality. Detailed information of our Endplate3D-QCT can be found in Table 1. Our dataset consists of a total of 170 scans, with each scan containing instance segmentation

annotations for 11 endplates. In other words, we provide a total of 1,870 annotated endplate instances, making this the largest publicly available endplate segmentation dataset to date.

Table 1. Summary statistics of the Endplate3D-QCT dataset. BMD (mg/cm^3): Bone Mineral Density measured in endplate region.

Metric	CTSpine-Refined	LDD-Mild	LDD-Severe	Overall
Image Count	119	25	26	170
Avg Image Size	(190, 188, 295)	(204, 209, 305)	(199, 202, 295)	-
Spacing (mm)	(0.8, 0.8, 0.8)	(0.4, 0.4, 0.75)	(0.39 0.39 0.74)	-
L1 endplate count	238	50	52	340
L1 endplate BMD	129.8 ± 6.8	128.8 ± 7.9	130.1 ± 7.0	129.7 ± 7.0
L2 endplate Count	238	50	52	340
L2 endplate BMD	129.0 ± 6.4	127.8 ± 7.2	129.9 ± 6.5	129.0 ± 6.6
L3 endplate Count	238	50	52	340
L3 endplate BMD	129.0 ± 5.9	127.6 ± 6.8	129.2 ± 5.7	128.8 ± 6.1
L4 endplate Count	238	50	52	340
L4 endplate BMD	128.9 ± 5.3	127.7 ± 5.8	128.4 ± 4.9	128.6 ± 5.3
L5 endplate Count	238	50	52	340
L5 endplate BMD	128.5 ± 6.1	128.3 ± 5.0	129.0 ± 4.3	128.6 ± 5.7
S1 endplate Count	119	25	26	170
S1 endplate BMD	128.4 ± 6.2	126.8 ± 4.7	127.8 ± 4.7	128.0 ± 7.3

2.2 Evaluation Protocol

Our automated evaluation framework employs multi-instance morphological analysis with clinical interpretability, formalized through the following mathematical constructs.

Volumetric Metrics Given a predicted segmentation mask P and ground truth G in 3D space $\Omega \subset \mathbb{Z}^3$, we define:

$$\mathrm{DSC}(P,G) = \frac{2|P \cap G|}{|P| + |G|} \in [0,1], \quad \mathrm{JI}(P,G) = \frac{|P \cap G|}{|P \cup G|} \in [0,1] \tag{1}$$

For surface distance metrics, let S_P and S_G denote the lumber endplate surfaces with point sets $\{p_i\}$ and $\{g_j\}$, respectively:

$$\mathrm{HD}_{95}(P,G) = \max\left\{ \underset{p \in S_P}{95\text{th}}\, d(p, S_G), \underset{g \in S_G}{95\text{th}}\, d(g, S_P) \right\} \tag{2}$$

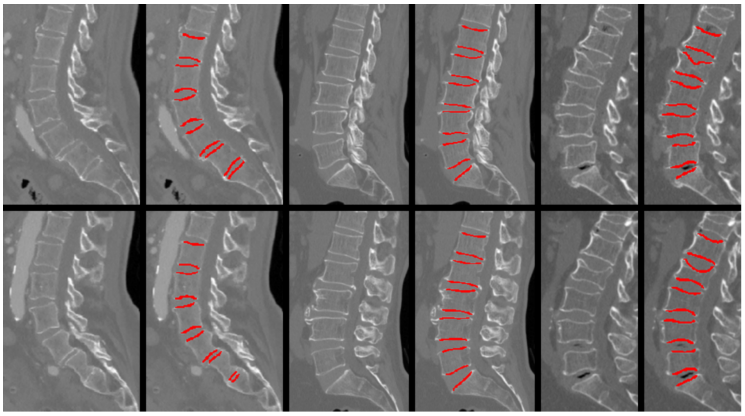

Fig. 1. Visualization of Endplate3D-QCT. The images in CTSpine-Refined were acquired from healthy individuals, with the alignment between vertebral bodies better conforming to the natural curvature of the human spine. In contrast, LDD-Mild exhibits a certain degree of degeneration, with some vertebrae showing signs of deformation. LDD-Severe represents more advanced degeneration, where some endplates have even undergone deterioration.

$$\mathrm{ASD}(P,G) = \frac{1}{|S_P| + |S_G|} \left(\sum_{i=1}^{|S_P|} d(p_i, S_G) + \sum_{j=1}^{|S_G|} d(g_j, S_P) \right) \tag{3}$$

Instance Matching For multi-instance endplate detection, define the optimal assignment matrix $A \in \{0,1\}^{M \times N}$ where M predicted instances and N ground truth instances:

$$A_{ij} = \begin{cases} 1, & \text{if } \arg\max_k \mathrm{DSC}(P_i, G_k) = j \text{ and } \mathrm{DSC}(P_i, G_j) > 0.8 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Precision, recall and F1 are then computed as:

$$\text{Precision} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} A_{ij}}{M}, \quad \text{Recall} = \frac{\sum_{j=1}^{N} \sum_{i=1}^{M} A_{ij}}{N} \tag{5}$$

and F1 is obtained through: $\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$.

3 Experiments

3.1 Implementation Details

We conducted experiments based on nnUNet, state-of-the-art training framework, and evaluate five widely-used methods, including UNet [11], VNet [9],

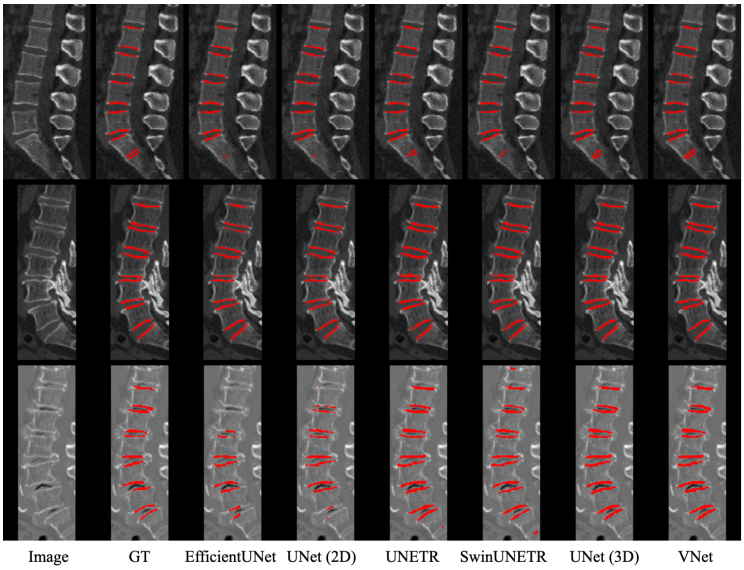

Fig. 2. Visualization of segmentation results. The first row presents a sample from CTSpine-Refined, where the models generally exhibit missed detections on the S1 endplate. The second row corresponds to LDD-Mild; due to the higher data resolution, the models generally achieve better performance. The third row shows examples from LDD-Severe, where severe endplate degeneration leads to poor model performance, particularly for EfficientNet and UNet (2D).

UNETR [5], SwinUNETR [4] and EfficientUNet [1]. All models were trained using Adam optimizer [8] with initial learning rate 3e-4 and cosine decay schedule. We implemented a hybrid loss function combining dice loss and cross entropy loss. For 3D methods, we used patch-based training and test-time sliding window inference. Data augmentation included random cropping, rotation, and scaling. Detailed implementation of each segmentation methods is shown in Table. 2. We performed three-fold cross-validation on the CTSpine-Refined subset and used an ensemble of the three models to make inferences on the other two subsets.

When evaluating each subset, we first calculate the instance-level Precision, Recall, and F1-score for endplates using Eq.5, which we refer to as the "Detection" metrics. Second, we compute the overall segmentation performance using Eq.1, which we define as the "Segmentation" metrics. Finally, we assess the segmentation performance for each individual endplate instance, including Dice, Jaccard, HD95, and ASD, which we categorize as the "Instance Segmentation" metrics.

3.2 Dataset-Centric Performance Analysis

Our experimental evaluation focuses on revealing how dataset characteristics influence segmentation performance across different anatomical presentations. We

Table 2. Implementation details of each method.

Model	Parameters (M)	Input Size	FLOPs (GFLOPs)
EfficientUNet [1]	10.05	(288, 288)	3.20
UNet (2D) [11]	18.67	(288, 288)	19.78
UNETR [5]	121.51	(320, 96, 96)	213.28
SwinUNETR [4]	15.7	(224, 96, 96)	184.01
UNet (3D) [11]	31.19	(320, 96, 96)	749.42
VNet [9]	9.46	(320, 96, 96)	178.85

Table 3. Performance comparison on *CTSpine-Refined* subset of our Endplate3D-QCT. The units of HD95 and ASD are millimeters (mm).

Method	Detection			Segmentation		Instance Segmentation			
	Precision	Recall	F1	Dice	Jaccard	Dice	Jaccard	HD95	ASD
EfficientUNet	0.904	0.909	0.906	0.867	0.766	0.878	0.785	1.007	0.258
UNet(2D)	0.945	0.942	0.943	0.885	0.795	0.891	0.804	1.003	0.225
UNETR	0.923	0.926	0.925	0.878	0.784	0.888	0.800	1.012	0.245
SwinUNETR	0.923	0.945	0.934	0.890	0.803	0.904	0.827	1.004	0.195
UNet(3D)	0.948	0.946	0.947	0.901	0.821	0.905	0.827	1.002	0.202
VNet	0.946	0.949	0.947	0.901	0.821	0.905	0.828	1.003	0.201

analyze three clinically distinct subsets: CTSpine-Refined (healthy morphology), LDD-Mild (early degeneration), and LDD-Severe (advanced pathology).

CTSpine-Refined: Baseline Performance. In cases with preserved anatomy (0.8mm isotropic voxels), all 3D methods achieved ≥ 0.92 F1 score for endplate detection. The uniform intensity profiles (HU range 1,200-1,500) enabled reliable boundary identification, with surface distance metrics showing submillimeter accuracy (ASD = 0.195–0.258mm).

However, even in this optimal scenario, as shown in Fig. 2, we observed false negatives in S1 endplate detection due to sacral curvature variations, highlighting inherent anatomical complexity.

LDD-Mild: Degenerative Onset. The 0.4mm axial resolution in early degeneration cases improved vertical structure visualization. Overall, the performance of different methods on the LDD-Mild subset is slightly better than on the CTSpine-Refined subset. This is mainly due to the higher resolution of images in LDD-Mild, which provides more detailed information. Additionally, the endplate degeneration in LDD-Mild is generally mild, posing less of a challenge for the segmentation models.

Table 4. Performance comparison on *LDD-Mild* subset of our Endplate3D-QCT. The units of HD95 and ASD are millimeters (mm).

Method	Detection			Segmentation		Instance Segmentation			
	Precision	Recall	F1	Dice	Jaccard	Dice	Jaccard	HD95	ASD
EfficientUNet	0.851	0.868	0.860	0.872	0.775	0.884	0.794	1.035	0.219
UNet (2D)	0.915	0.938	0.926	0.887	0.797	0.891	0.805	1.008	0.225
UNETR	0.947	0.947	0.947	0.895	0.810	0.898	0.816	1.018	0.233
SwinUNETR	0.946	0.949	0.947	0.905	0.827	0.909	0.835	1.014	0.201
UNet (3D)	0.972	0.961	0.966	0.915	0.844	0.916	0.847	1.005	0.187
VNet	0.960	0.956	0.958	0.918	0.849	0.919	0.852	1.010	0.187

Table 5. Performance comparison on *LDD-Severe* subset of our Endplate3D-QCT. The units of HD95 and ASD are millimeters (mm).

Method	Detection			Segmentation		Instance Segmentation			
	Precision	Recall	F1	Dice	Jaccard	Dice	Jaccard	HD95	ASD
EfficientUNet	0.600	0.554	0.576	0.773	0.641	0.862	0.759	1.123	0.232
UNet (2D)	0.704	0.728	0.716	0.832	0.716	0.869	0.770	1.139	0.273
UNETR	0.778	0.785	0.782	0.853	0.747	0.880	0.788	1.139	0.294
SwinUNETR	0.803	0.818	0.811	0.863	0.761	0.886	0.798	1.120	0.261
UNet (3D)	0.887	0.891	0.889	0.880	0.787	0.895	0.813	1.112	0.258
VNet	0.875	0.870	0.872	0.885	0.796	0.900	0.821	1.125	0.253

LDD-Severe: Pathological Extremes. Severely degenerated cases (0.39mm spacing, extensive osteophytes) exposed fundamental limitations. Despite having the same voxel dimensions as LDD-Mild, the increased pathological complexity led to a decline in all evaluation metrics, highlighting the challenges of LDD-Severe. Although the Segmentation performance did not drop significantly (e.g., Dice decreased from 0.918 to 0.885 for VNet), the decline in Detection metrics was particularly pronounced, with Recall dropping below 0.9 across all methods. Notably, EfficientUNet had the lowest Recall, reaching only 0.554. We believe that the decline in performance of all models on this subset is primarily due to the excessive morphological variability of the endplates in these cases, resulting in their irregularity.

4 Conclusion

In this study, we introduce the open-access dataset Endplate3D-QCT, featuring pixel-level annotations of lumbar endplates in clinical quantitative CT (QCT) scans. Building upon this dataset, we developed an automated evaluation framework for robust performance assessment. Extensive experiments were conducted using widely-used segmentation models to perform comprehensive benchmark-

ing. These resources establish a foundation for advancing endplate-specific analysis, contributing to improved research in spinal biomechanics, personalized implant design, and bone density mapping.

While our dataset represents a significant advancement in endplate characterization, current deep learning methods still fall short of the precision required for clinical application. Despite achieving high Dice scores on relatively healthy spines, all tested models exhibit inconsistencies in identifying endplates, particularly in the presence of severe degenerative changes. To bridge the gap between research and clinical applicability, future efforts should develop more robust segmentation techniques capable of handling pathological variations. Furthermore, intraoperative validation studies are essential to establish direct correlations between segmentation accuracy and surgical outcomes, ensuring that AI-assisted methods meet the stringent requirements of clinical decision-making.

Acknowledgments. This research was supported in part by the Brain-like General Vision Model and Applications project (2022ZD0160403).

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

References

1. Baheti, B., Innani, S., Gajre, S., Talbar, S.: Eff-unet: A novel architecture for semantic segmentation in unstructured environment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 358–359 (2020)
2. Deng, Y., Wang, C., Hui, Y., Li, Q., Li, J., Luo, S., Sun, M., Quan, Q., Yang, S., Hao, Y., et al.: Ctspine1k: A large-scale dataset for spinal vertebrae segmentation in computed tomography. arXiv preprint arXiv:2105.14711 (2021)
3. van der Graaf, J.W., van Hooff, M.L., Buckens, C.F., Rutten, M., van Susante, J.L., Kroeze, R.J., de Kleuver, M., van Ginneken, B., Lessmann, N.: Lumbar spine segmentation in mr images: a dataset and a public benchmark. Scientific Data **11**(1), 264 (2024)
4. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images (2022), https://arxiv.org/abs/2201.01266
5. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.: Unetr: Transformers for 3d medical image segmentation (2021), https://arxiv.org/abs/2103.10504
6. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
7. Juanying, X., Ying, P., Mingzhao, W.: The squeeze & excitation normalization based nnu-net for segmenting head & neck tumors. Chinese Journal of Electronics **33**(3), 766–775 (2024). https://doi.org/10.23919/cje.2022.00.306, https://cje.ejournal.org.cn/en/article/doi/10.23919/cje.2022.00.306
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

9. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. arXiv preprint arXiv:1606.04797 (2016)
10. Möller, H., Graf, R., Schmitt, J., Keinert, B., Schön, H., Atad, M., Sekuboyina, A., Streckenbach, F., Kofler, F., Kroencke, T., et al.: Spineps—automatic whole spine segmentation of t2-weighted mr images using a two-phase approach to multi-class semantic and instance segmentation. European Radiology pp. 1–12 (2024)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. arXiv preprint arXiv:1505.04597 (2015)
12. Zhang, Z., Liu, T., Fan, G., Li, B., Feng, Q., Zhou, S.: Spinemamba: Enhancing 3d spinal segmentation in clinical imaging through residual visual mamba layers and shape priors. arXiv preprint arXiv:2408.15887 (2024)