

PO-IoU 기반 영점추론 무릎 관절 분할 기법

이재훈^{1o}

유재천^{1*}

¹성균관대학교 전기전자컴퓨터공학과

{jachoonlee, yoojc}@skku.edu

PO-IoU: Point-Offset-Guided Zero-Shot Knee Joint Segmentation for Osteoarthritis Assessment

Jaehoon Lee^{1o}

Jaechern Yoo^{1*}

¹Department of Electrical and Computer Engineering, Sungkyunkwan University

요약

무릎 관절 X-ray 영상에서의 자동 분할은 골관절염 분석의 대규모 자동화를 위해 중요하지만, 픽셀 단위 분할 정답 주석을 구축하는 데 큰 비용이 든다. 본 논문에서는 few-shot 관절 위치 검출, zero-shot 분할, 포인트 오프셋 기반 프롬프트 보정을 결합한, 픽셀 단위 분할 정답 없이 동작하는 무릎 관절 분할 프레임워크인 PO-IoU를 제안한다. YOLO 기반 검출기로 관절 관심영역을 추출하고 MedSAM으로 후보 마스크를 생성한 뒤, 전체 데이터셋 수준에서 히스토그램 기반 구조 프록시와의 중첩을 최대화하는 최적 프롬프트 위치를 도출하여 분할 결과를 안정화한다. 8,252장의 방사선 영상 실험에서 PO-IoU는 Kellgren–Lawrence 등급과 유의한 음의 상관관계를 보였으며, 등급별 평균 추세에서 $R^2 = 0.819$ 를 나타냈다.

Abstract

Automated segmentation of the knee joint region in X-ray images is important for large-scale automation of osteoarthritis (OA) analysis, but constructing pixel-wise segmentation ground-truth annotations is costly. We propose PO-IoU, a knee joint segmentation framework that operates without pixel-wise segmentation ground truth by combining few-shot joint localization, zero-shot segmentation, and point-offset prompt calibration. A YOLO-based detector extracts the joint region of interest and MedSAM generates candidate masks; the optimal prompt position that maximizes overlap with a histogram-derived structural proxy is then derived at the dataset level to stabilize the segmentation result. In experiments on 8,252 radiographs, PO-IoU showed a significant negative correlation with the Kellgren–Lawrence grade, with $R^2 = 0.819$ for the grade-wise mean trend.

키워드: 무릎 골관절염, MedSAM, 포인트 오프셋 보정, 영점추론 분할, 의료영상 정보학

Keywords: Knee osteoarthritis, MedSAM, point-offset calibration, zero-shot segmentation, medical image informatics

1 Introduction

Knee osteoarthritis (OA) is a prevalent degenerative disease in which progressive cartilage loss narrows the radiographic joint space [1]. In routine practice, joint-space morphology on weight-bearing X-rays is used to assess disease severity through the Kellgren–Lawrence (KL) grading system [1, 2]. Manual assessment remains common but is labor-intensive and prone to inter-

observer variability, particularly in large longitudinal cohorts [3, 4].

Although deep learning has improved automated analysis of knee radiographs [5, 6, 7], two barriers remain important in data-scarce settings. First, high-performing segmentation models typically depend on dense pixel-wise annotations [8, 9]. Second, zero-shot foundation models such as MedSAM are sensitive to prompt placement and can leak into adjacent bony structures when bound-

*corresponding author: Jaechern Yoo / Department of Electrical and Computer Engineering, Sungkyunkwan University (yoojc@skku.edu)

aries are weak or degraded [10, 11]. These limitations motivate a pipeline that reduces annotation burden while stabilizing zero-shot segmentation.

We therefore formulate knee joint segmentation as an integrated three-stage pipeline: few-shot JROI localization, zero-shot MedSAM segmentation, and PO-IoU-based prompt calibration. The proposed method uses histogram-derived structural information inside the detected JROI to identify the population-level prompt offset that yields the most consistent masks. Every stage is deployable at modest annotation cost: sparse box annotations are sufficient for localization, zero-shot MedSAM provides candidate masks, and the final selection step is driven by image statistics. This makes the overall framework appealing for retrospective cohort analysis where curated segmentation masks are unavailable but radiographs and severity labels are abundant.

The main contributions of this study are threefold. First, we present PO-IoU, a prompt-calibration criterion that scores MedSAM candidate masks against a histogram-derived structural proxy in place of unavailable pixel-wise ground truth. The design is deliberately minimal—an off-the-shelf detector, a zero-shot segmenter, and a single population-level offset—so that the entire pipeline can be transferred to a new cohort without any segmentation label, additional training, or per-image tuning. Second, we combine few-shot YOLO localization and zero-shot prompt-based segmentation into a single end-to-end practical workflow for knee radiographs. Third, we show that the resulting PO-IoU values remain clinically informative at scale by exhibiting a significant negative relationship with KL grade in a large OA cohort.

2 Related Work

Before deep learning, knee joint-space assessment relied on semi-automated and classical computer-vision methods. Early approaches applied Canny edge detection and morphological operations to locate the tibia–femur boundary, from which joint-space width (JSW) was estimated along a series of vertical profiles [32, 33]. Active shape and active appearance models later enabled more flexible bone contour extraction [34], and atlas-based registration methods aligned a template skeleton to each new radiograph to derive JSW measurements automatically [35]. While these methods achieve clinically acceptable precision in controlled settings, they require careful parameter tuning, depend on consistent acquisition geometry, and typically need expert review before deployment on large heterogeneous cohorts. Their reliance on precise bone boundary localization also makes them sensitive to the same low-contrast and sclerotic conditions that challenge deep learning approaches. The proposed pipeline sidesteps explicit bone con-

tour estimation entirely: by anchoring MedSAM to the radiolucent joint-space interior via a histogram-derived proxy, PO-IoU operates on intensity statistics rather than boundary geometry, offering a complementary route that is less sensitive to boundary degradation.

Recent studies have shown that deep neural networks are effective for knee OA analysis from plain radiographs, including severity grading, ROI localization, and structural assessment [5, 12, 13]. Yet many of these methods rely on carefully curated labels or dense annotations that are difficult to scale in clinical practice. Follow-up studies have explored few-shot learning, external validation, and broader review perspectives, underscoring both the promise and the remaining generalization gap of current OA systems [14, 15, 16, 7]. Detection-oriented pipelines have also improved structural focus by combining localized analysis with classification or grading [17, 18, 19].

Foundation segmentation models such as SAM and MedSAM offer a promising alternative by enabling prompt-based zero-shot inference [20, 10, 21]. In medical imaging, however, prompt sensitivity remains a major limitation: small changes in prompt position can substantially alter the predicted mask, especially under low contrast and weak anatomical boundaries [11, 22, 23]. In practice, many prompt-automation approaches reduce user burden but still depend on task-specific heuristics, additional training signals, or domain adaptation, which can weaken portability across cohorts. Our work addresses this limitation by replacing manual prompt selection with a histogram-guided objective tailored to knee radiographs. By grounding prompt selection in the local brightness distribution of the target image itself, the proposed method preserves the zero-shot nature of MedSAM while introducing a simple anatomical prior tied to the radiolucent joint space. This distinction is important for clinical translation: instead of retraining a new segmentation model for every institutional dataset, the refinement logic can be reused as long as the local radiographic appearance of the joint space remains diagnostically meaningful.

3 Proposed Method

The proposed framework consists of three stages, as summarized in Fig. 1: (1) image preprocessing and few-shot JROI localization, (2) zero-shot MedSAM segmentation, and (3) PO-IoU-based prompt calibration.

In the first stage, each input radiograph undergoes Z-score normalization followed by CLAHE (Contrast Limited Adaptive Histogram Equalization, clip limit 0.01) to reduce acquisition-dependent intensity variation and improve local contrast around the knee joint [24, 25]. This preprocessing standardizes the visual con-

through PO-IoU calibration. Let μ_{JROI} and σ_{JROI} denote the mean and standard deviation of pixel intensities within the detected JROI. We define an adaptive threshold as

$$\tau = \mu_{\text{JROI}} - 0.5\sigma_{\text{JROI}},$$

and binarize pixels below τ to construct the proxy region Ω_{hist} . This proxy approximates the radiolucent joint space by exploiting the fact that the joint space tends to occupy the darker local region after contrast normalization, adapting naturally to patient-specific intensity ranges and acquisition conditions (Fig. 2).

Prompt candidates are generated through a two-stage grid search centered on the JROI. Let $\Omega_{\text{SAM}}(p)$ denote the SAM mask produced from prompt $p = (c_x + \varepsilon, c_y + \theta)$. The PO-IoU score for each candidate is defined as

$$J(p) = \frac{|\Omega_{\text{SAM}}(p) \cap \Omega_{\text{hist}}|}{|\Omega_{\text{SAM}}(p) \cup \Omega_{\text{hist}}|}.$$

In Stage 1 (coarse search), horizontal offsets $\varepsilon \in [-20, +20]$ and vertical offsets θ spanning both upward and downward directions are evaluated at a coarse step to map the full optimization landscape. The resulting mean PO-IoU heatmap reveals that warm-color (high-scoring) regions concentrate consistently in the downward half of the search space, while upward shifts ($\theta < 0$) yield uniformly lower scores. This observation is anatomically consistent: the radiolucent joint space lies in the lower portion of the detected bounding box, between the femoral condyle above and the tibial plateau below. Motivated by this finding, Stage 2 (refined downward search) restricts the vertical range to $\theta \in [0, +30]$ pixels and sweeps $\varepsilon \in [-20, +20]$ at step size 2 px, yielding a 21×16 candidate grid. Fig. 3 visualizes both search stages: the coarse landscape in panel (a) and the refined landscape with the identified global maximum in panel (b).

The grid search is conducted once across the full dataset. The global maximum identified in Stage 2, ($\hat{\varepsilon} = 0, \hat{\theta} = 10$), is then applied as a fixed bias at every inference, making deployment deterministic and eliminating per-image grid search overhead. After MedSAM inference, connected-component analysis retains only the largest foreground region to suppress noise blobs. PO-IoU thus serves as a prompt calibration criterion: rather than requiring a manually placed prompt, the framework identifies the population-level offset at which the joint space most consistently aligns with the histogram proxy, and fixes that offset for all subsequent inference.

4 Experimental Results

All experiments were run on a single workstation with an NVIDIA GeForce RTX 3090 GPU (24 GB) under Python 3.11 and PyTorch 2.5.1 with CUDA 12.1. Detection used Ultralytics YOLOv8n (v8.3.235), and zero-shot segmentation used the ViT-B backbone [10, 11]; image I/O and processing used OpenCV 4.12 and NumPy 1.26. Because the pipeline performs inference only—no per-image optimization is carried out and a single population-level offset is applied at test time—the procedure is deterministic and reproducible given the same inputs and model weights.

Experiments were conducted on 8,252 knee radiographs annotated with KL grades. The radiographs were drawn from a publicly available knee osteoarthritis grading dataset [36], derived from the Osteoarthritis Initiative (OAI) repository and comprising frontal knee X-rays graded KL 0–4, where left and right knees are treated as independent images. The grade distribution is markedly imbalanced—approximately 3,252, 1,494, 2,174, 1,085, and 250 images for KL 0 through KL 4, respectively—reflecting the natural prevalence of mild-to-moderate disease. Because the segmentation stage is zero-shot and performs no per-image training, the evaluation does not depend on a patient-level train/test partition; the only trained component is the lightweight YOLO localizer, fit from 52 bounding-box annotations. Public knee OA grading resources and cohort-oriented evaluation settings provide useful context for interpreting this type of large-scale analysis [29, 4]. Instead of evaluating direct geometric measurements, we examined how PO-IoU varies with disease severity. This evaluation strategy is motivated by the fact that projection-dependent geometric measurements can fluctuate with positioning and acquisition setup, whereas a structure-consistency score can still provide useful population-level information even when individual measurements are noisy. Linear regression on individual images ($N = 8,252$) revealed a statistically significant negative association between PO-IoU and KL grade ($p < 0.001$, $R^2 = 0.048$). The low sample-level R^2 is expected given large image-quality variation, class imbalance, and anatomical diversity. A separate regression on the per-grade mean values showed a clear decreasing trend with $R^2 = 0.819$, as illustrated in Fig. 4. Because a regression fitted to five aggregated points carries limited information, we characterize the same association at the individual-image level using several independent tests, summarized in Table 2. Rank correlations, a non-parametric comparison across the five grades with its effect size, and the discrimination between early and advanced disease all confirm the negative trend on the full cohort; the grade-level regression is reported alongside them for reference rather than as primary evidence. Because the KL grade never enters the calibration objective, which is defined

the same quantity is used both to calibrate the prompt offset and to summarize the resulting masks; it should accordingly be read as a consistency measure rather than as a substitute for accuracy metrics such as Dice or Hausdorff distance. Two aspects of the design delimit this dependence: the KL grade used for validation does not enter the calibration objective, which is defined solely on the overlap with Ω_{hist} , and the calibration is applied as a single population-level bias rather than optimized per image. The criterion remains meaningful within an anatomically plausible range of prompt displacements; establishing accuracy in the conventional sense nonetheless requires expert pixel-level annotation.

Scope of comparison. Quantitative comparison against classical joint-space-width pipelines, supervised segmentation networks, or alternative prompt-automation strategies requires pixel-level ground truth on a common subset; the configuration comparison in Fig. 5 is accordingly illustrative in nature.

Level of aggregation. The association with KL grade is robust in aggregate ($H = 635.4$, $p = 3.4 \times 10^{-136}$) while individual-image variation remains large ($R^2 = 0.048$). The score is therefore best used as a cohort-level indicator of structural degradation rather than as a per-patient measurement.

Partition, cohort, and proxy assumptions. The calibration offset ($\hat{\varepsilon}$, $\hat{\theta}$) was determined on the same cohort on which results are reported; left and right knees are treated as independent images; and all radiographs derive from a single publicly available OAI-derived source, leaving robustness across acquisition protocols untested. The proxy threshold assumes that the joint space forms the darkest local region after normalization, which may not hold under severe sclerosis, radiopaque implants, or end-stage ankylosis. Finally, the framework operates on 2D frontal radiographs and models neither bilateral correspondence, longitudinal change, nor 3D modalities such as MRI.

These considerations define our ongoing work. We are assembling an expert-annotated subset with pixel-level masks stratified across KL grades, which will allow the pipeline and its configurations to be scored with Dice, IoU, and 95% Hausdorff distance and compared directly against classical and supervised baselines. Extending the criterion with explicit plausibility terms—containment within the detected joint region, single-connected-component structure, and an elongation prior consistent with joint-space morphology—would allow the offset to be optimized rather than bounded. Beyond a single global offset, a lightweight per-image predictor fitted under a patient-level train/validation/test partition, together with validation on an independent cohort and on the longitudinal structure of the OAI cohort, would establish whether the calibration transfers and whether the score tracks joint-space narrowing within subjects over time. The same proxy-guided cal-

ibration principle may also transfer to other low-contrast musculoskeletal segmentation tasks where annotation cost and prompt sensitivity pose similar challenges.

References

- [1] J. H. Kellgren and J. S. Lawrence, "Radiological assessment of osteoarthritis," *Annals of the Rheumatic Diseases*, vol. 16, no. 4, pp. 494–502, 1957.
- [2] M. D. Kohn, A. A. Sassoon, and N. D. Fernando, "Classifications in brief: Kellgren-Lawrence classification of osteoarthritis," *Clinical Orthopaedics and Related Research*, vol. 474, no. 8, pp. 1886–1893, 2016.
- [3] L. Sharma, "Osteoarthritis of the knee," *New England Journal of Medicine*, vol. 384, no. 1, pp. 51–59, 2021.
- [4] C. G. Peterfy, E. Schneider, and M. Nevitt, "The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee," *Osteoarthritis and Cartilage*, vol. 16, no. 12, pp. 1433–1441, 2008.
- [5] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach," *Scientific Reports*, vol. 8, no. 1, p. 1727, 2018.
- [6] P. Phan-Trung, T. Tran-Trung, and H. Nguyen, "Knee osteoarthritis detection and severity grading using YOLO-based architecture," in *Proc. Int. Conf. on Computing and Communication Technologies (RIVF)*, pp. 1–6, 2023.
- [7] Y. Zhao, Z. Li, and T. Liu, "Multi-scale feature fusion for knee osteoarthritis grading in X-ray images," *Medical Image Analysis*, vol. 99, p. 103352, 2025.
- [8] B. Norman, V. Padoia, and S. Majumdar, "Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry," *Radiology*, vol. 288, no. 1, pp. 177–185, 2018.
- [9] A. Swiecicki, N. Li, J. O'Donnell, W. Said, D. Zhang, B. Roth, and J. E. Burdick, "Deep learning-based algorithm for assessment of knee osteoarthritis severity in radiographs matches performance of radiologists," *Computers in Biology and Medicine*, vol. 133, p. 104334, 2021.
- [10] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.
- [11] Y. Huang, Z. Yang, H. Liu, J. Zhou, and X. Han, "Segment anything model for medical images?," *Medical Image Analysis*, vol. 92, p. 103061, 2024.
- [12] G. Wang, J. E. Peek, and A. Ross, "Automated measurement of joint space width in hand radiographs for assessment of osteoarthritis," *Computerized Medical Imaging and Graphics*, vol. 93, p. 101970, 2021.
- [13] M. Yunus, H. Latif, Z. Gao, and A. Yousaf, "Automated knee joint bone area segmentation using deep learning," *IEEE Access*, vol. 10, pp. 70513–70525, 2022.
- [14] C. P. Littlefield, M. J. Trochez, and S. M. Bhatt, "External validation of deep learning models for knee osteoarthritis severity grading," *Osteoarthritis and Cartilage Open*, vol. 5, no. 2, p. 100363, 2023.
- [15] E. Vaattovaara, A. Tiulpin, and S. Saarakkala, "Longitudinal deep learning analysis of knee osteoarthritis progression," *npj Digital Medicine*, vol. 8, no. 1, p. 21, 2025.
- [16] P. Rani, R. Singh, and A. Kumar, "A comprehensive review of deep learning approaches for knee osteoarthritis analysis from radiographs," *Artificial Intelligence in Medicine*, vol. 160, p. 102950, 2026.
- [17] N. Pongsakonpruttikul, C. Lohwongwatana, and P. Sirikulchayanonta, "Automated knee joint localization and grading in X-ray images using convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 75, p. 103582, 2022.
- [18] T. Abdullah, T. Ali, and S. Hussain, "Knee osteoarthritis severity grading based on deep learning and X-ray images," *Diagnostics*, vol. 12, no. 3, p. 597, 2022.
- [19] J. Antony, K. McGuinness, K. Moran, and N. E. O'Connor, "Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks," in *Proc. Int. Conf. on Machine Learning and Data Mining (MLDM)*, pp. 376–390, 2017.
- [20] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pp. 4015–4026, 2023.
- [21] H. Wang, S. Guo, J. Ye, Z. Deng, J. Cheng, T. Li, J. Chen, Y. Su, Z. Huang, Y. Shen, and others, "SAM-Med3D: towards general-purpose segmentation models for volumetric medical images," *arXiv preprint arXiv:2310.15161*, 2023.
- [22] Y. Zhang and J. Jiao, "How segment anything model (SAM) boost medical image segmentation?," *arXiv preprint arXiv:2305.03678*, 2023.
- [23] J. Sun, J. Li, Y. Pan, J. Wu, and Z. Wen, "AutoSAM: adapting SAM to medical images by overcoming catastrophic forgetting," *arXiv preprint arXiv:2306.13740*, 2024.
- [24] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, and K. E. Muller, "Contrast-limited adaptive histogram equalization: speed and effectiveness," in *Proc. IEEE Conf. on Visualization in Biomedical Computing*, pp. 337–345, 1990.
- [25] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *Journal of VLSI Signal Processing Systems*, vol. 38, no. 1, pp. 35–44, 2004.
- [26] N. Bayramoglu, M. Nieminen, S. Saarakkala, and J. Tiulpin, "Self-supervised learning for joint localization and automatic grading in knee X-rays," *Medical Image Analysis*, vol. 64, p. 101724, 2020.
- [27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.
- [28] Tzutalin, "LabelImg," *GitHub repository*, 2015. [Online]. Available: <https://github.com/tzutalin/labelImg>
- [29] P. Chen, L. Gao, X. Shi, K. Allen, and L. Yang, "Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 84–92, 2019.
- [30] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, 2019.
- [31] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports*, vol. 12, no. 1, p. 5979, 2022.

- [32] J. Duryea, J. C. Buckland-Wright, E. Keen, C. Hayball, R. Marshall, P. Rogers, and P. Shipley, "A fully automated software approach to the measurement of joint space width in hand radiographs," *Medical Physics*, vol. 27, no. 4, pp. 869–877, 2000.
- [33] J. A. Lynch, S. Parimi, C. E. Chaganti, N. E. Lane, and M. C. Nevitt, "The association of proximal femoral shape and incident radiographic hip OA in elderly women," *Osteoarthritis and Cartilage*, vol. 18, no. 1, pp. 38–44, 2010.
- [34] F. Ding and A. Leahy, "Automated joint space width measurement and articular surface contouring in knee radiographs using active appearance models," *Journal of Medical Imaging*, vol. 1, no. 1, p. 014001, 2014.
- [35] E. B. Dam, M. Lillholm, J. Marques, and M. Nielsen, "Automatic segmentation of high- and low-field knee MRIs using knee image quantification with data from the OA initiative," *Journal of Medical Imaging*, vol. 2, no. 2, p. 024001, 2015.
- [36] P. Chen, "Knee osteoarthritis severity grading dataset," *Mendeley Data*, V1, 2018. doi: 10.17632/56rmx5bjcr.1.

〈 저 자 소 개 〉



유 재 천

- Current Position: Professor
- Affiliation: Department of Electrical and Computer Engineering, Sungkyunkwan University
- Education
- Ph.D. in Electrical and Electronic Engineering, POSTECH, 2001
- Career / Experience
- Senior Research Engineer, Samsung Electronics Co., Ltd., 1986 - 1996
- Professor, Samsung Institute of Technology, 1994
- Adjunct Professor, POSTECH Information Research Laboratories and Graduate School for Information Technology, 2002 - 2006
- Research Assistant Professor, POSTECH Graduate School for Information Technology, 2006 - 2008
- 관심분야: Design of deep learning architectures and their applications, AI-based medical image diagnosis, Design of AI-controlled robotic hands, AI-based digital healthcare platforms
- <https://orcid.org/0000-0003-3721-0207>



이 재 훈

- Current Position: Master's Student
- Affiliation: Department of Electrical and Computer Engineering, Sungkyunkwan University
- Education
- B.S. in Computer Engineering, Kyungdong University, 2024
- M.S. Student in Department of Electrical and Computer Engineering, Sungkyunkwan University, 2025 - Present
- 관심분야: Image processing, Medical AI
- <https://orcid.org/0009-0000-5448-1848>