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.

OrthoScribe: Training-Free Multimodal Orthodontic Report Generation via Photographic Retrieval and 3D Geometry Correction

Sanket Kachole[1,2] and Spyridon Bakas[1,2,3,4]*

[1] Division of Computational Pathology, Department of Pathology and Laboratory Medicine, Indiana University School of Medicine (IUSM), Indianapolis, IN, USA
[2] IU Melvin and Bren Simon Comprehensive Cancer Center, Indianapolis, IN, USA
[3] Departments of Biostatistics and Health Data Science; Radiology and Imaging Sciences; Neurological Surgery; IUSM, Indianapolis, IN, USA
[4] Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University, Indianapolis, IN, USA
† Corresponding author: spbakas@iu.edu

Abstract. Malocclusion is a orthodontic condition of teeth misalignment, between the upper and lower jaws, affecting roughly half of the global population. Each orthodontic case must be documented in a structured diagnostic report describing occlusal relationships and dental findings. Automatic generation of orthodontic reports from intraoral data may reduce documentation burden and improve consistency. Orthodontic assessment is inherently multimodal: occlusal relationships are best derived from 3D scans, whereas soft-tissue and restorative findings are primarily visible in intraoral photographs. Here, we introduce OrthoScribe, a training-free computational approach that retrieves a similar case report using self-supervised image features, removes unverifiable patient-specific claims, and corrects occlusal descriptors using measurements from the 3D meshes. The method was developed on the ODIN 2026 Bite2Text dataset, comprising 997 cases, of which 869 provided usable paired scans, photographs, and English reports, split into 781 training and 88 validation cases. On internal validation, OrthoScribe achieved a RadFact logical F1 of 0.363, outperforming a constant-report baseline (0.326), and a fine-tuned vision–language model (0.274). On the official held-out multicenter test phase, OrthoScribe achieved BLEU-4 and METEOR scores of 0.173 and 0.388, respectively. Code and trained weights are available on GitHub and Zenodo.

Keywords: Report generation · Multimodal learning · Retrieval · Orthodontics · Intraoral imaging.

1 Introduction

Malocclusion, or "*bad bite*" is an orthodontic condition of teeth misalignment, describing the relationship between the upper and lower jaws. It affects roughly half

* Corresponding author: spbakas@iu.edu

of the global population, with a large share of children and adolescents requiring orthodontic assessment [24]. Each malocclusion case must be documented in a structured diagnostic report including occlusal relationships and dental findings. Such assessment integrates dental casts, intraoral photographs, and clinical evaluation of occlusion to produce reports that guide treatment planning. Automated visual analysis has been used to reduce manual assessment burden in monitoring applications [13], while recent advances in vision–language models have enabled report generation with dedicated factuality evaluation [1, 3]. Language models are also increasingly used to integrate structured clinical measurements with free-text descriptions beyond radiology [5]. Orthodontic reporting remains comparatively underexplored despite the substantial global burden of oral disease and the need for improved clinical documentation [23].

The MICCAI 2026 Oral and Dental Image aNalysis (ODIN) Bite2Text challenge [17] addresses malocclusions by pairing 3D intraoral scans, intraoral photographs, and expert reports, that provide complementary information, requiring modality-aware processing rather than uniform fusion [20,25,27], a problem also observed in combining event and frame data [12]. Similar benefits of modality-aware integration have been reported in multimodal cancer diagnosis and prognosis [16,26]. Factual faithfulness is also critical, as incorrect tooth references or occlusal descriptors are more consequential than stylistic variation [6,18].

To address the ODIN Bite2Text challenge, we introduce OrthoScribe, a training-free computational reporting approach that separates photographic evidence from measurable geometry. Query photographs are encoded using a self-supervised image model [21], and the report of a geometrically re-ranked similar training case is retrieved as an initial draft. A precision filter removes unverifiable patient-specific claims, including explicit tooth numbers and tooth-specific restorations, while a geometry module derives overjet and overbite from the registered upper and lower meshes and corrects the corresponding report descriptors. For comparison, an open vision–language model [1] was fine-tuned using low-rank adaptation [8], together with several training and retrieval variants. OrthoScribe requires no task-specific training, satisfies the ODIN Bite2Text challenge compute constraints, and produces coherent reports for evaluation. Unlike existing retrieval-based report generators [7], which retrieve text within a single image modality, the novelty here is a cross-modal division of labor: photographs drive retrieval, whereas an independent 3D measurement channel both re-ranks candidates and deterministically overwrites the occlusal descriptors it can measure, so that a single correction simultaneously removes an unsupported claim and restores the correct one. These efficiency considerations are consistent with the motivation for lightweight neural architectures in event-based vision [15].

2 Materials and Methods

2.1 Materials

The ODIN 2026 Bite2Text dataset [17] contains 997 cases, each comprising 3D intraoral scans of the upper and lower arches, intraoral photographs, and ex-

Table 1. Dataset composition, split, and report statistics. Modality rows justify using only photograph reports as targets.

Property	Value
Total cases released	997
Cases with English photograph report	872
Scan-only cases (unusable target)	125
Dropped (missing report or mesh)	4
Usable paired cases	**869**
Training / validation split	781 / 88
Report length (median words / sentences)	118 / 9
Unique token vocabulary	1,483
Intraoral photographs (total)	5,192
Photographs per case (typical / max)	5 / 9
Median photograph resolution	3972×2501
Visual keywords: photo vs. scan reports	92% vs. 3%
Photo/scan report text overlap	0.3%
Geometry-derivable share of claims	$\approx$0.61
Recall ceiling (occlusal only)	$\leq$0.61

Table 2. Keyword-derived descriptor prevalence over 872 photograph reports. Occlusal descriptors (left) form a near-universal skeleton; photographic findings (right) are variable. Teeth are named in 75% of reports (mean 4.6; 218 name none, 139 name ten or more).

Occlusal descriptor	%	Photographic finding	%
Molar relationship	99.9	Gingival findings	79.8
Canine relationship	99.8	Restorations	58.6
Curve of Spee	99.9	Sealants / fissures	43.0
Curve of Wilson	99.9	Plaque / calculus	31.9
Vertical (overbite)	99.5	Edentulous / missing	31.0
Midline	99.2	Recession	25.6
Overjet	98.5	Staining / pigment.	23.7
Transverse	98.3	Caries	21.9
Crowding (low./up.)	73.7 / 69.8	Prosthetic crowns	6.5

pert orthodontic reports. A dataset audit identified cases suitable for report-generation experiments and characterized the information required for faithful reporting, consistent with established practices for curating and documenting datasets involving emerging sensing modalities [9]. Among the released cases, 872 contained English photograph-based reports and 125 contained scan-only reports. The latter were excluded as they largely omitted restorative and soft-tissue findings visible in the photographs. Visual-finding keywords occurred in 92% of photograph reports compared with 3% of scan reports, with only 0.3% text overlap between the two report types. After excluding four additional cases

lacking a usable report or mesh, 869 paired cases remained and were divided into 781 training and 88 validation cases (Table 1). Reports contained a median of 118 words across nine sentences and followed a consistent structure covering sagittal, vertical, and transverse relationships, midlines, curves of Spee and Wilson, crowding, and photographic findings. Claim-proxy analysis indicated that approximately 61% of reference claims were geometry-derivable, imposing an approximate recall ceiling of 0.61 on a geometry-only system and demonstrating the importance of photographic evidence for the remaining findings. Keyword-based prevalence analysis (Table 2) showed that core occlusal descriptors occurred in nearly all reports, forming a stable reporting skeleton, whereas photographic findings were substantially more variable. Explicit tooth references appeared in 75% of reports, with a mean of 4.6 teeth named per report, identifying patient-specific tooth-level statements as a major source of potential false claims and motivating the precision-filtering stage of the proposed pipeline.

2.2 Methods

Fig. 1 shows the OrthoScribe workflow. The system separates the two modalities: the intraoral photographs drive retrieval of a similar report, while the 3D meshes provide exact occlusal measurements used to correct that report. The following subsections describe each stage.

Preprocessing. Photographs were converted to a common color format and resized such that the shorter image dimension matched the encoder input size; truncated or corrupted files were handled defensively. No data augmentation was applied, although augmentation strategies have demonstrated improved robustness in other sensing domains [19]. The registered upper and lower 3D meshes were loaded in a format-agnostic manner and subsampled to a fixed number of vertices for computational efficiency. Graph-based and neuromorphic approaches provide alternative strategies for processing sparse geometric representations directly [10, 11]. As the scans were provided in arbitrary orientations, a canonicalization procedure established consistent left–right, anterior–posterior, and vertical axes from the paired arches before geometric measurements were derived.

Retrieval of a report draft. Each intraoral photograph was encoded using a self-supervised vision transformer [21], and case-level representations were obtained by averaging the embeddings across all available photographs. For each query, the $k = 5$ training cases with the highest cosine similarity were retained as candidates. Because intraoral photographs are visually homogeneous and cosine similarities among the top candidates were near-identical, the final draft was selected by re-ranking these candidates on occlusal geometry. The candidate minimizing the standardized Euclidean distance in overjet and overbite to the query meshes was chosen, and its report was used as the initial draft. This

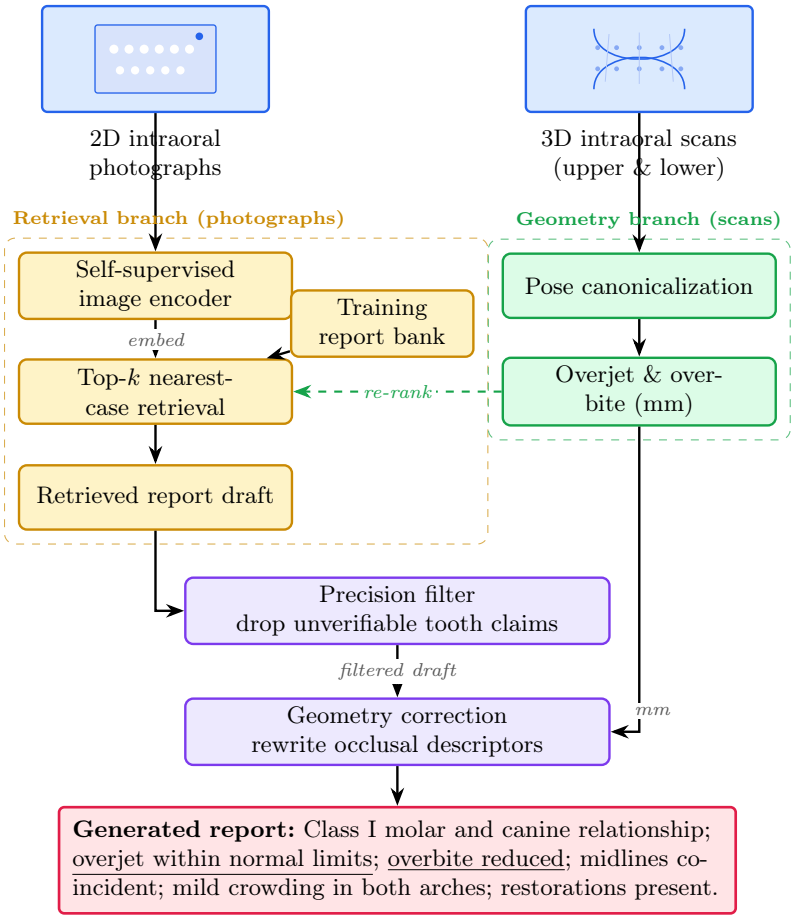

Fig. 1. The OrthoScribe approach workflow, shown top to bottom. The intraoral photographs (left, amber) are encoded with a self-supervised image model and matched against a bank of training reports to retrieve a draft. The 3D scans (right, green) are canonicalized and measured to obtain overjet and overbite, which additionally re-rank the retrieval candidates so that the selected draft matches the query occlusion (dashed arrow). A precision filter then removes patient-specific claims that cannot be verified, and a geometry-correction step rewrites the occlusal descriptors using the measured values (underlined in the example), yielding the final report.

strategy exploits the consistent structure of orthodontic reports while retaining common occlusal statements, providing a coherent starting point without free-form generation.

Precision filtering. Retrieved reports originate from different patients and may therefore contain unsupported patient-specific findings, particularly explicit

tooth references and restorations assigned to individual teeth. The precision filter removes these tooth-number references and restoration-to-tooth assignments while preserving the general occlusal description. This conservative filtering reduces unverifiable claims while retaining clinically relevant report content.

Geometry correction. The canonical frame was obtained by axis selection rather than decomposition. The vertical axis was taken as the direction of smallest arch extent, confirmed by the largest upper–lower centroid separation. The mediolateral axis as the remaining direction of greatest bilateral mirror symmetry. The anteroposterior axis as the last direction, oriented so that the narrower arch end faced anteriorly. Within this frame, the incisal edge of each arch was taken as the most anterior vertices near the midline, refined by a lateral band median and an anteroposterior percentile to suppress outliers. Overjet and overbite were then measured as the horizontal and vertical separations between the upper and lower incisal points. Continuous values were mapped by fixed thresholds, i.e., overjet below 0 mm reduced and above 4.5 mm increased, whereas overbite below 0.5 mm reduced and above 4.5 mm increased, and otherwise within normal limits. These measurements provided the most reliable geometric signals, and the corresponding descriptors in the retrieved draft were replaced with values consistent with the query anatomy. Direct derivation of anatomical descriptors from sensor geometry has similarly demonstrated reliability in other applications, including posture estimation from tactile signals [14]. Correcting an erroneous occlusal descriptor simultaneously removes a false claim and restores the corresponding correct claim, thereby improving both report precision and recall.

Vision–language baseline. For comparison, Qwen2.5-VL-7B-Instruct [1] was fine-tuned on the training reports using LoRA [8], with up to four intraoral photographs at 448-pixel shorter-side resolution and a textual summary of the geometric measurements provided as input. The input resolution was constrained by the deployment memory budget. Adapters of rank 32 and scaling 64 were applied to all attention and feed-forward projections with the vision encoder frozen, and training used a constant learning rate of 2×10^{-4} with warmup for ten epochs at an effective batch size of eight in bfloat16, selecting the final checkpoint by validation loss. Greedy decoding without repetition penalties was adopted, as stronger decoding constraints destabilized generation. The resulting model served as the generative baseline for comparison with the retrieval-based pipeline. Related work has similarly investigated conditioning generative models on clinical attributes, including pathology-conditioned gait synthesis [4].

2.3 Evaluation

Reports were evaluated using RadFact logical F1 [3] for bidirectional claim-level entailment and BLEU-4 [22] and METEOR [2] for lexical overlap. The overall score was $S = 0.8 \times \text{RadFact} + 0.1 \times \text{BLEU-4} + 0.1 \times \text{METEOR}$. RadFact was

Table 3. RadFact evaluation on the 88-case internal validation split using the challenge judge model. Precision and recall are the claim-level components of RadFact logical F1. Experiments are grouped by vision–language model and retrieval-based development, with "+" indicating incremental modifications to the preceding base method. The best value in each column is shown in bold.

Experiment	Precision	Recall	RadFact F1
Baseline and Vision–Language Model Development			
Medoid Constant-Report Baseline	0.379	0.286	0.326
LoRA VLM Baseline	0.339	0.230	0.274
+ Multilingual Supervision	0.337	0.306	0.320
+ Expanded Adapter Capacity	0.272	0.165	0.205
+ Structured-Template Generation	0.178	0.085	0.115
Retrieval-Based Development			
Nearest-Neighbor Report Retrieval	0.313	0.325	0.319
+ Claim Precision Filtering	0.366	0.310	0.336
+ **Geometry Correction (OrthoScribe)**	**0.406**	**0.328**	**0.363**

computed on the 88-case validation set; BLEU-4 and METEOR came from the official hidden-test leaderboard. Attribute-level accuracies (Section 3) quantify per-descriptor correctness. Test Phase factuality could not be assessed because the leaderboard returned only lexical-overlap metrics.

3 Results and Discussion

Table 3 summarizes internal RadFact development. The constant-report and LoRA VLM baselines achieved F1 scores of 0.326 and 0.274, respectively. The latter produced incorrect tooth-level claims. An unquantified short cosine-decay schedule underfit, while constant-rate training resolved the generation failure. Multilingual supervision improved F1 to 0.320, but expanded adapter capacity (0.205) and structured-template generation (0.115) degraded it, indicating that capacity and output constraints did not resolve factual grounding. Nearest-neighbor retrieval achieved an F1 of 0.319 and recall of 0.325. Precision filtering removed unverifiable tooth and restoration claims, raising precision from 0.313 to 0.366 and F1 to 0.336. Geometry correction produced the final *OrthoScribe* system, with precision 0.406, recall 0.328, and F1 0.363. Thus, photographic retrieval supplied report structure, filtering removed unsupported claims, and 3D geometry corrected measurable occlusal findings.

Official leaderboard results are reported separately in Table 4. On the fully held-out multicenter Test Phase, OrthoScribe obtained BLEU-4 and METEOR scores of 0.173 ± 0.099 and 0.388 ± 0.109, respectively. The lower absolute scores relative to internal validation indicate the greater difficulty of generalizing to unseen centers and acquisition conditions.

The experiments support our final design. The VLM could not reliably localize tooth-specific restorations at the deployable resolution, making it less faithful

Table 4. Official challenge leaderboard performance of the final OrthoScribe system. Scores are reported as mean ± standard deviation across cases.

Evaluation Phase	BLEU-4	METEOR
Debugging Phase	0.437 ± 0.402	0.585 ± 0.296
Test Phase (held-out multicenter)	0.173 ± 0.099	0.388 ± 0.109

than retrieval and filtering. Geometry was useful for correction but not retrieval because similar occlusal measurements did not imply similar photographic findings. Only overjet (0.756 vs. 0.372 majority baseline) and overbite (0.480 vs. 0.320) exceeded their baselines, whereas transverse, curve of Wilson, curve of Spee, and midline did not. We therefore restricted correction to overjet and overbite and inherited other descriptors from the retrieved draft. Combining these corrections with removal of low-confidence patient-specific claims improved factual precision without sacrificing recall.

4 Conclusion

We developed OrthoScribe, a training-free computational approach that searches through a cohort of orthodontic reports, paired with intraoral photographs, and retrieves the one whose photographs are "closest" to the currently assessed case. It then removes unverifiable patient-specific claims from the retrieved report, and corrects overjet and overbite statements using the 3D intraoral scan of the assessed case. It outperformed constant-report and fine-tuned vision–language baselines in validation factuality, while official Test Phase evaluation was limited to lexical-overlap metrics. Its main limitations are dependence on the report bank, correction of only two geometric descriptors, and unavailable claim-level or expert Test Phase evaluation. Future work will extend geometric correction and incorporate orthodontist assessment.

Acknowledgments. This work was partially supported by the Lilly Endowment, Inc., through the Indiana University Pervasive Technology Institute. The content is solely the authors' responsibility and does not represent the official views of the funding body.

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

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005)

3. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: MAIRA-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
4. Chandrasekaran, M., Kachole, S., Francik, J., Makris, D.: PGcGAN: Pathological gait-conditioned gan for human gait synthesis. arXiv preprint arXiv:2603.14409 (2026)
5. Chandrasekaran, M., Kachole, S., Francik, J., Makris, D.: LLM-conditioned synthesis of pathological gaits via structured gait-language representations. arXiv preprint arXiv:2606.06048 (2026)
6. Delbrouck, J.B., Chambon, P., Bluethgen, C., Tsai, E., Almusa, O., Langlotz, C.: Improving the factual correctness of radiology report generation with semantic rewards. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 4348–4360 (2022)
7. Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., Rajpurkar, P.: Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In: Machine learning for health. pp. 209–219. PMLR (2021)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
9. Huang, X., Kachole, S., Ayyad, A., Naeini, F.B., Makris, D., Zweiri, Y.: A neuromorphic dataset for tabletop object segmentation in indoor cluttered environment. Scientific data **11**(1), 127 (2024)
10. Kachole, S.: Object Segmentation: From Neuromorphic Sensing to Neuromorphic Machine Learning. PhD thesis, Kingston University London (2024)
11. Kachole, S., Alkendi, Y., Naeini, F.B., Makris, D., Zweiri, Y.: Asynchronous events-based panoptic segmentation using graph mixer neural network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4083–4092 (2023)
12. Kachole, S., Huang, X., Naeini, F.B., Muthusamy, R., Makris, D., Zweiri, Y.: Bimodal SegNet: Fused instance segmentation using events and rgb frames. Pattern Recognition **149**, 110215 (2024)
13. Kachole, S., Hunter, G.J., Duran, O.: A computer vision approach to monitoring the activity and well-being of honeybees. In: Intelligent Environments (Workshops). pp. 152–161 (2020)
14. Kachole, S., Nayak, B., Brouner, J., Liu, Y., Guo, L., Makris, D.: Posture estimation from tactile signals using a masked forward diffusion model. Sensors **25**(16), 4926 (2025)
15. Kachole, S., Sajwani, H., Naeini, F.B., Makris, D., Zweiri, Y.: Asynchronous bioplausible neuron for spiking neural networks for event-based vision. In: European Conference on Computer Vision. pp. 399–415. Springer (2024)
16. Kachole, S., Thakur, S., Innani, S., Adap, S., You, S., Pitarch-Abaigar, C., Bakas, S.: Multi-FRuGaL: Multimodal flexible redundancy-aware decomposed gated learning for cancer diagnosis and prognosis. arXiv preprint arXiv:2606.06867 (2026)
17. Lumetti, L., Rizzo, F., Cremonini, F., Candeloro, E., Luca, L., Grana, C., Bolelli, F., et al.: Do multimodal llms understand intraoral dental data? dataset, platform, and baselines. In: Computer Vision–ECCV 2026 (2026)
18. Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5288–5304 (2021)

19. Naeini, F.B., Kachole, S., Muthusamy, R., Makris, D., Zweiri, Y.: Event augmentation for contact force measurements. IEEE Access **10**, 123651–123660 (2022)
20. Noeldeke, B., Vassis, S., Sefidroodi, M., Pauwels, R., Stoustrup, P.: Comparison of deep learning models to detect crossbites on 2d intraoral photographs. Head & Face Medicine **20**(1), 45 (2024)
21. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
23. Rehman, U., Hudson, H., Hao, C.Y., Ahn, Y., Kachole, S., Liu, J., Patel, S., Xu, M.J., Rouhani, M.J., O'Flynn, P., et al.: Lifetime prevalence of betel nut chewing in india and taiwan: Raising awareness of oral cancer risks and the urgent call for regulation. Cancers **18**(7), 1074 (2026)
24. Stomatologic, S.: Worldwide prevalence of malocclusion in the different stages of dentition: A systematic review and meta-analysis. Eur J Paediatr Dent **21**(2), 115–22 (2020)
25. Wu, T.H., Lian, C., Lee, S., Pastewait, M., Piers, C., Liu, J., Wang, F., Wang, L., Chiu, C.Y., Wang, W., et al.: Two-stage mesh deep learning for automated tooth segmentation and landmark localization on 3d intraoral scans. IEEE transactions on medical imaging **41**(11), 3158–3166 (2022)
26. You, S., Pitarch-Abaigar, C., Kachole, S., Sonawane, S., Ha, J., Gada, A.S., Crandall, D., Shiradkar, R., Bakas, S.: PROFUSEme: PROstate cancer biochemical recurrence prediction via FUSEd multi-modal embeddings. arXiv preprint arXiv:2509.14051 (2025)
27. Zanjani, F.G., Moin, D.A., Verheij, B., Claessen, F., Cherici, T., Tan, T., et al.: Deep learning approach to semantic segmentation in 3d point cloud intra-oral scans of teeth. In: International Conference on Medical Imaging with Deep Learning. pp. 557–571. PMLR (2019)