Semi-Supervised Dental CBCT Report Generation via Entity-Aware Dual-Space Retrieval

Lele Han[1]* and Xinshuo Ma[1]

School of Cyberspace Security, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China
HanLele@hdu.edu.cn
maxinshuo421@163.com

Abstract. Dental cone-beam computed tomography (CBCT) provides detailed three-dimensional views of teeth and surrounding structures, but converting these volumes into structured reports remains difficult. Accurate reporting requires joint anatomical reasoning and consistent identification of tooth-level findings, diagnoses, treatments, and medications. Limited paired supervision further complicates the learning of reliable image-to-report associations. We present an extractive, entity-aware framework for structured dental CBCT report generation. The framework first represents each volume with frozen multiscale VoCo-B features and explicit anatomy descriptors. It then uses anti-curriculum pseudo-labelling (ACPL) to train a compact multi-label predictor, incorporating image-only samples according to their informativeness and pseudo-label reliability. At inference, predicted entities guide field-specific retrieval from observed reports, while deterministic report-memory priors provide content for fields that CBCT cannot support reliably. This design separates image-grounded findings from fields that depend more heavily on recurring report patterns. In the official evaluation, our method achieved a Weighted Score of 0.2364 and generated a prediction for every case.

Keywords: Dental CBCT · Report Generation · Semi-Supervision · Entity Prediction · Report Retrieval

1 Introduction

Dental cone-beam computed tomography (CBCT) shows teeth and surrounding maxillofacial structures in three dimensions [3], preserving density patterns and spatial relationships that planar radiographs collapse into a single plane. The reporting task considered here converts these volumes into seven-field dental reports that preserve tooth notation, diagnosis codes, treatment actions, and medication terms. This task requires more than fluent text generation. A report must connect local image evidence to the correct clinical entities and place those

* Corresponding author.

entities in the appropriate fields. A plausible sentence is still unreliable when it introduces findings or treatments that the scan does not support. The main difficulty therefore lies in preserving both anatomical correspondence and entity consistency throughout the report.

Our design draws on three related areas: three-dimensional representation learning, semi-supervised learning, and medical report generation. Swin UNETR extracts hierarchical features from volumetric images [6], while VoCo pretrains volumetric encoders through volume contrastive learning [18]. Semi-supervised medical image analysis has used selective mutual learning, consistency regularization, uncertainty-guided mixing, and deep co-training to learn from images without manual annotations [5, 7, 10, 17, 19, 20]. ACPL focuses pseudo-label training on informative samples in low-density regions [9]. Mean Teacher reduces short-term prediction fluctuations through an exponential moving average [14], and Asymmetric Loss handles the imbalance between positive and negative labels [13]. Medical report systems take either a generative or retrieval-based approach. R2Gen [2] and R2GenCMN [1] use memory-driven decoding, MedWriter retrieves reports and sentences [21], KERP combines retrieval with paraphrasing [8], and CT2Rep generates reports from three-dimensional chest CT [4]. These methods provide useful components, but dental CBCT reporting still needs tooth-level entity learning and tighter control over the content assigned to each report field.

We build an *entity-aware field-routing* framework around this requirement. The framework follows three connected stages. The representation stage combines five frozen VoCo-B feature scales with an explicit anatomy descriptor and projects them into compact spaces, retaining learned volumetric context alongside structural cues. The entity-learning stage trains a compact multi-label predictor with ACPL. A Mean Teacher, neighbour propagation, and reliability weighting construct soft targets for informative image-only samples, while Asymmetric Loss handles label imbalance. Once entity learning is complete, the grounded-synthesis stage retrieves report text at the field level. Predicted entities re-rank observed candidates for image-supported fields, whereas deterministic report-memory priors provide content for fields with weak CBCT evidence. The synthesis stage restricts every output string to text already present in the report memory. In the official evaluation, the framework achieved a Weighted Score of 0.2364 and returned predictions for every query. Since the official challenge provides no matched baseline, Table 3 reports this score as an absolute result.

2 Methods

2.1 Overview

Indices i, j, and q denote paired, image-only, and query examples. Let $x_i \in \mathbb{R}^{W\times H\times D}$ be a CBCT volume with spatial dimensions W, H, D. Its report is $r_i = \{r_i^f\}_{f\in\mathcal{F}}$, where r_i^f is field f's text and $\mathcal{F}$ contains the seven fields. The entity vector $y_i \in \{0,1\}^C$ covers C categories; $y_{ic} = 1$ means category c occurs in report i. The paired and image-only sets are $\mathcal{D}_l = \{(x_i, r_i, y_i)\}$ and $\mathcal{D}_u = \{x_j\}$.

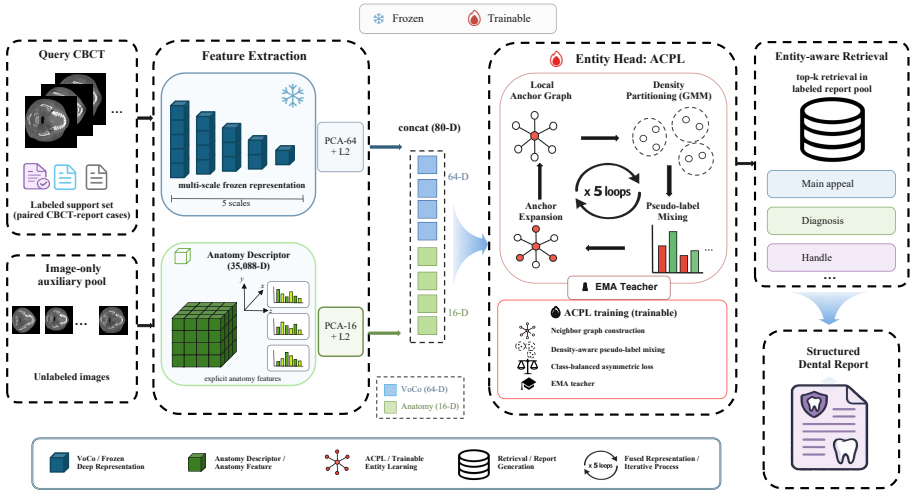


Fig. 1. Overview of the framework. Frozen multiscale VoCo-B features and anatomy descriptors form two compact spaces. ACPL trains the entity head through local anchors, density-based sample selection, neighbour propagation, and an exponential-moving-average teacher. Predicted entities then guide field-specific retrieval from the report memory. A deterministic memory prior supplies fields with weak CBCT evidence, and every output string remains extractive.

For field f, the candidate index set and its distinct text memory are

$$\mathcal{I}_f = \{i : r_i^f \neq \varnothing\}, \qquad \mathcal{T}_f = \text{unique}\{r_i^f : i \in \mathcal{I}_f\}.$$

Thus, $i \in \mathcal{I}_f$ means that field f is non-empty for case i, and $\mathcal{T}_f$ is its distinct observed text. For query q, $p_q \in [0,1]^C$ gives entity probabilities and $\widehat{r}_q^f \in \mathcal{T}_f$ denotes the predicted field text.

Figure 1 summarizes the method. The representation module maps each CBCT to separate semantic and anatomy spaces. The entity module concatenates the two projections and learns a multi-label predictor with reliability-weighted ACPL. The retrieval module combines image similarity and entity evidence to select field-level text. The following subsections describe these modules in the same order.

2.2 Dual-space CBCT representation

We resample each volume to a common grid $m = (m_W, m_H, m_D)$ while preserving its physical field of view. For axis $\nu \in \{W, H, D\}$, let m_ν, n_ν, and s_ν denote the target grid size, original grid size, and original voxel spacing. The target spacing $\widetilde{s}_\nu$ is

$$\widetilde{s}_\nu = s_\nu \frac{\max(n_\nu - 1, 1)}{\max(m_\nu - 1, 1)}.$$

Linear interpolation retains the original origin and direction. We clip each volume between its lower and upper intensity percentiles and normalize the result to $[0, 1]$.

We denote the clipped and normalized volume by $\overline{x}_i$. The semantic branch passes it through a frozen VoCo-B encoder. Let $H_s(\overline{x}_i)$ be the output of encoder stage s, for $s = 1, \ldots, 5$. Global average pooling (GAP) over these five outputs produces

$$v_i = \operatorname{concat}_{s=1}^{5} \operatorname{GAP}(H_s(\overline{x}_i)) .$$

Thus, v_i is the concatenated semantic descriptor. The anatomy branch forms descriptor a_i from pooled intensity maps, high-density maps, three-axis projections, and an intensity histogram. Separate principal component analysis (PCA) projections preserve the geometry of each branch:

$$z_i^v = \operatorname{norm}(P_v \operatorname{norm}(v_i)) , \qquad z_i^a = \operatorname{norm}(P_a \operatorname{norm}(a_i)) .$$

Here, P_v, P_a denote the fitted PCA matrices, $\operatorname{norm}(\cdot)$ is L2 normalization, and z_i^v, z_i^a are the 64- and 16-dimensional semantic and anatomy projections. Their concatenation $z_i = [z_i^v; z_i^a]$ forms the 80-dimensional entity input, while retrieval keeps the two spaces separate.

2.3 Reliability-weighted entity learning

The entity vocabulary covers tooth notation, diagnosis codes and terms, treatment actions, and medication terms. The entity head h_θ, with trainable parameters θ, maps z_i to $p_i = \sigma(h_\theta(z_i))$, where σ is the elementwise sigmoid. The head contains LayerNorm, a linear hidden layer, GELU, dropout, and a linear output layer. We denote its L2-normalized hidden embedding by e_i; these embeddings define the ACPL neighbourhood graph.

Supervised learning uses Asymmetric Loss [13]. For category c, where $c = 1, \ldots, C$, let N_c^+, N_c^- be its positive and negative counts and α_c its positive weight:

$$\alpha_c = \operatorname{clip}\left(\sqrt{\frac{N_c^- + 1}{N_c^+ + 1}}, 1, 6\right) , \tag{1}$$

$$\ell(y_{ic}, p_{ic}) = -y_{ic}\alpha_c \log p_{ic} - (1 - y_{ic}) p_{ic}^2 \log(1 - p_{ic}).$$

Here, $\ell(y_{ic}, p_{ic})$ is the loss for target y_{ic} and probability p_{ic}; it upweights sparse positives and downweights easy negatives.

ACPL maintains a student with parameters θ_S and a teacher with parameters θ_T. At update iteration t, the teacher follows the student through

$$\theta_T \leftarrow \mu_t \theta_T + (1 - \mu_t)\theta_S, \qquad \mu_t = \min\left(0.99, 1 - \frac{1}{t + 2}\right) .$$

where μ_t is the exponential-moving-average coefficient. For image-only embedding e_i, $\mathcal{N}_k(i)$ contains its $k = 5$ nearest labelled or accepted pseudo-labelled

anchors, which produce

$$y_i^{\mathrm{nbr}} = \sum_{j \in \mathcal{N}_k(i)} \frac{\exp(\mathrm{sim}(e_i, e_j)/\tau)}{\sum_{m \in \mathcal{N}_k(i)} \exp(\mathrm{sim}(e_i, e_m)/\tau)}\, y_j, \qquad \tau = 0.10,$$

where sim is cosine similarity, τ is the softmax temperature, and y_j is the target of anchor j. The mean neighbour similarity and normalized local density are

$$\bar{s}_i = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \mathrm{sim}(e_i, e_j), \qquad d_i = \mathrm{clip}\left(\frac{\bar{s}_i - s_{.02}}{s_{.98} - s_{.02}}, 0, 1 \right),$$

where $s_{.02}, s_{.98}$ are the 2nd and 98th percentiles of $\bar{s}_i$ over image-only samples; $d_i = 0.5$ when they coincide. A three-component Gaussian mixture partitions samples by $\bar{s}_i$, and ACPL expands anchors from the lowest-density component. The density d_i mixes teacher probabilities q_i with the neighbour target:

$$\widetilde{y}_i = d_i q_i + (1 - d_i) y_i^{\mathrm{nbr}}.$$

For entity category c, define binary entropy $H(u) = -u \log u - (1-u) \log(1-u)$. Teacher–neighbour agreement A_i and pseudo-label confidence C_i are

$$A_i = 1 - \frac{1}{C} \sum_{c=1}^{C} |q_{ic} - y_{ic}^{\mathrm{nbr}}|, \qquad C_i = 1 - \frac{1}{C \log 2} \sum_{c=1}^{C} H(\widetilde{y}_{ic}).$$

They define reliability ρ_i and pseudo-label loss weight w_i:

$$\rho_i = \mathrm{clip}(0.55 A_i + 0.45 C_i, 0.1, 1), \qquad w_i = 0.45 \rho_i.$$

The objective sums ℓ over categories and adds $w_i \ell(\widetilde{y}_i, p_i)$ for selected pseudo-labelled samples; w_i affects entity learning only.

2.4 Entity-aware field retrieval

For query index q and candidate index $j \in \mathcal{I}_f$, let s_{qj}^v and s_{qj}^a denote cosine similarities in the semantic and anatomy spaces. Their rank-normalized combination gives

$$B_{qj} = 0.55 R(s_{qj}^v) + 0.45 R(s_{qj}^a),$$

where $R(\cdot)$ maps descending candidate ranks linearly to $[0, 1]$, and B_{qj} is the dual-space image score. Query probabilities p_q and candidate entities y_j define field-specific compatibility:

$$\begin{aligned} E_{qj}^f &= \begin{cases} 1, & d_{qj}^f < 10^{-8}, \\ \dfrac{2 \sum_{c \in \mathcal{C}_f} p_{qc} y_{jc}}{d_{qj}^f}, & \text{otherwise}, \end{cases} \\ d_{qj}^f &= \sum_{c \in \mathcal{C}_f} p_{qc} + \sum_{c \in \mathcal{C}_f} y_{jc}, \\ S_{qj}^f &= B_{qj} + \beta_f R(E_{qj}^f), \end{aligned} \tag{2}$$

where $\mathcal{C}_f$ contains field-relevant entity indices, E^f_{qj} is their soft F1 compatibility, and S^f_{qj} is the final candidate score. The zero-denominator convention assigns full compatibility when neither vector contains a field-relevant entity. The coefficient β_f controls entity evidence; we set it to 0.50, 1.00, and 0.25 for Main appeal, Diagnosis, and Handle, respectively.

For Main appeal, Diagnosis, and Handle, the router takes the top-k candidates under S^f_{qj}. Repeated texts receive frequency priority; otherwise, the router returns

$$\mathrm{medoid}(T) = \arg\max_{t \in T} \sum_{\substack{t' \in T \\ t' \neq t}} J_{\mathrm{tok}}(t, t'),$$

where T is the multiset of texts from the top-ranked candidates and J_{tok} is token-set Jaccard similarity. Length and lexicographic order resolve ties.

The remaining fields use the deterministic memory prior

$$\pi_f = \arg\max_{t \in \mathcal{T}_f} \sum_{\substack{t' \in \mathcal{T}_f \\ t' \neq t}} J_{\mathrm{tok}}(t, t').$$

Here, π_f is the medoid of the complete text memory for field f. Let $\mathcal{K}^f_q$ denote the top-k candidate indices under S^f_{qj}. The routing rule is

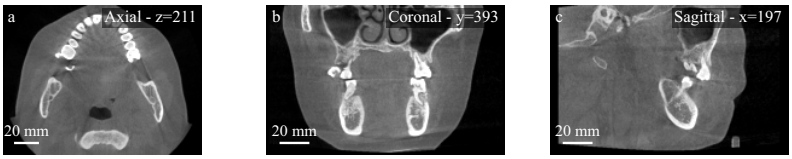
$$\widehat{r}^f_q = \begin{cases} \mathrm{medoid}\Big(\{r^f_j : j \in \mathcal{K}^f_q\}\Big), & f \in \{\text{Main appeal, Diagnosis, Handle}\}, \\ \pi_f, & \text{otherwise.} \end{cases}$$

The first branch keeps image-supported fields query dependent; the second assigns a shared representative to fields with weak CBCT evidence. Every selected string comes from $\mathcal{T}_f$, so the method does not generate new text.

3 Experiments

3.1 Dataset and strictly isolated protocol

The official STSR 2026 Task 3 dataset comprises 300 mutually disjoint CBCT cases: 50 paired Train-Labeled cases, 200 image-only training cases, and 50 Validation cases whose reference reports are hidden [11]. The paired cases contain seven-field reports and provide supervised entity targets and the only reports used to construct the report memory. All 200 image-only cases are used to fit the semantic and anatomy PCA mappings and form the pseudo-labelled pool for ACPL entity learning; because they have no reports, they never enter the report memory or field-specific retrieval candidates. Validation CBCTs are inference queries only and contribute to no PCA fitting, ACPL training, report-memory construction, model selection, hyperparameter tuning, or calibration. We freeze every component before Validation inference. Figure 2 pairs three orthogonal views of case 515 with its seven-field prediction. Ellipses shorten long fields, and the hidden reference precludes case-level assessment.



d Model-generated structured report — Validation case 515 - frozen-pipeline output

Hidden reference unavailable; this panel shows inference output, not case-level correctness.

Field	Model inference text
Main appeal	Pain in upper right back tooth for several days
Present medical history	The patient reported that his lower front teeth had been loose for several years and now he came to our hospital for treatment. The patient reported that the ...
Oral Check	36 The teeth are missing, the alveolus is slightly sunken, and the overall oral health condition is acceptable. The gums in the 46 implantation area healed ...
Diagnosis	Chronic apical periodontitis (K04.500) Marginal gingivitis (K05.104).
Treatment plan	16 26 Crown restoration is recommended after root canal treatment.
Handle	Caries removal, pulp opening, false wall made with glass ionomer, 3 tubes explored, 20 mm proximal to tongue, 18 mm proximal to buccal, (mesial 2 in 1) 19 mm ...
Doctor advices	Follow-up for discomfort.

Fig. 2. Illustrative inference for Validation case 515. (a–c) Orthogonal CBCT views. (d) Frozen-pipeline seven-field output, shortened with ellipses. The hidden reference precludes case-level assessment. The image panels contain no segmentation overlay.

3.2 Official evaluation protocol

The organizer-provided Validation scorer supplies the formal performance evidence. It compares seven-field reports with hidden references and returns the Weighted Score, language-overlap metrics, entity precision, recall and F1, diagnosis matching, and field completion. We report only the frozen full method because no matched official baseline or ablation output is available.

3.3 Implementation and training protocols

Each run used one 48-GB NVIDIA RTX A6000 GPU. We initialized the frozen VoCo-B encoder from its public checkpoint [18], cached its features once, and optimized the 41,186-parameter ACPL entity head. Table 1 lists the hardware and software.

Training used full-batch updates on cached features with Gaussian noise ($\sigma = 0.012$) and no voxel-space augmentation. Model selection used only Train-Labeled.

4 Results and Discussion

4.1 Official evaluation completes all 50 Validation cases

The official scorer returned *Success* for all 50 Validation cases and a Weighted Score of 0.2364. Predicted and ground-truth field completion were 1.0000 and 0.9000. Table 3 reports the score and entity F1.

Table 1. Development environment.

Item	Configuration	Item	Configuration
System	Linux; distribution not recorded	CUDA / cuDNN	12.1 / 8.9.2
CPU	2× Xeon Gold 6326	Language	Python 3.10.19
RAM	256 GB	Framework	PyTorch 2.1.2; MONAI 1.3.0
GPU	2× RTX A6000, 48 GB; one used per run		

Table 2. Training, complexity, and runtime protocol.

Item	Configuration
Pre-trained model	Public VoCo-B; encoder frozen
Input / batch	$1 \times 64 \times 128 \times 128$; feature batch 4
Augmentation	No voxel augmentation; feature noise $\sigma = 0.012$
Sampling	Full-batch updates; low-density GMM; top-$k = 5$
Steps / fit	450 warm-up $+5 \times 220$ pseudo-label steps
Optimizer / LR	AdamW; 3×10^{-3}, then 1.5×10^{-3}
Loss / regularization	Reliability-weighted asymmetric loss; weight decay 2×10^{-3}; clip 5.0
Model selection	Two seeds × five folds; final 50-case fit
Training times	∼103.5 s for selection and final fit; ∼138.2 s for the cached pipeline
Parameters	72.8033 M instantiated; 18.6785 M active; 0.0412 M trainable
FLOPs	65.4766 G/query, active path (`fvcore`)
Memory / inference	6.59 GiB; 50 cases in 246 s (4.92 s/case)

Table 3. Official performance of the full method on 50 Validation cases.

Method	W. score	Tooth	Diagn.	Treat.	Med.
Entity-aware dual-space retrieval (full method)	0.2364	0.1304	0.0940	0.3226	0.1867

Treatment actions had the highest F1 (0.3226), whereas diagnosis exact matching was 0.0400. Without a matched baseline, stage-level attribution and comparative claims are unavailable.

Treatment-action, tooth, and treatment-plan recall were 0.6467, 0.4230, and 0.7067; diagnosis partial and exact matching were 0.1300 and 0.0400. The 1.0000 prediction-completion rate measures coverage, not patient-level correctness.

4.2 Limitations and evidence boundary

The paired cohort limits entity and report-memory diversity, and extractive selection does not guarantee case-level factuality. Memory priors provide query-independent text where image evidence is weak; they do not establish patient-specific correctness. Clinical claims require de-identified expert assessment.

5 Conclusion

The entity-aware framework combines multiscale VoCo-B and anatomy features with reliability-weighted entity learning and extractive field retrieval. It achieved a Weighted Score of 0.2364 and returned all required fields for every query. This is an absolute benchmark result, not a comparative or clinical claim; future

work should add query-specific evidence, examine diagnosis errors, and assess factuality through expert review.

Acknowledgements This work was developed and evaluated within the ODIN 2026 challenge cluster. We thank the organizers and data providers of the associated challenge programmes [16, 15, 11, 12], including STS, ToothFairy, and Bite2Text.

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

References

1. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 5904–5914 (2021). https://doi.org/10.18653/v1/2021.acl-long.459
2. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. pp. 1439–1449 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.112
3. Cui, Z., Li, C., Wang, W.: ToothNet: Automatic tooth instance segmentation and identification from cone beam CT images. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6368–6377 (2019). https://doi.org/10.1109/CVPR.2019.00653
4. Hamamci, I.E., Er, S., Menze, B.: CT2Rep: Automated radiology report generation for 3D medical imaging. In: Medical Image Computing and Computer-Assisted Intervention. pp. 476–486. Lecture Notes in Computer Science (2024). https://doi.org/10.1007/978-3-031-72390-2_45
5. Hang, W., Dai, P., Pan, C., Liang, S., Zhang, Q., Wu, Q., Jin, Y., Wang, Q., Qin, J.: Pseudo-label guided selective mutual learning for semi-supervised 3D medical image segmentation. Biomedical Signal Processing and Control **100**, 107144 (2025). https://doi.org/10.1016/j.bspc.2024.107144
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Crimi, A., Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Lecture Notes in Computer Science, vol. 12962, pp. 272–284 (2022). https://doi.org/10.1007/978-3-031-08999-2_22
7. Lei, T., Liu, H., Wan, Y., Li, C., Xia, Y., Nandi, A.K.: Shape-guided dual consistency semi-supervised learning framework for 3-D medical image segmentation. IEEE Transactions on Radiation and Plasma Medical Sciences **7**(7), 719–731 (2023). https://doi.org/10.1109/TRPMS.2023.3286866
8. Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6666–6673 (2019). https://doi.org/10.1609/aaai.v33i01.33016666

9. Liu, F., Tian, Y., Chen, Y., Liu, Y., Belagiannis, V., Carneiro, G.: ACPL: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20665–20674 (2022). https://doi.org/10.1109/CVPR52688.2022.02004
10. Liu, Y., Tian, Y., Wang, C., Chen, Y., Liu, F., Belagiannis, V., Carneiro, G.: Translation consistent semi-supervised segmentation for 3D medical images. IEEE Transactions on Medical Imaging **44**(2), 952–968 (2025). https://doi.org/10.1109/TMI.2024.3468896
11. Lumetti, L., Di Bartolomeo, M., Pellacani, A., Anesi, A., Grana, C., Bolelli, F.: Ontology-grounded structured prediction for dental CBCT reporting. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2026 (2026)
12. Lumetti, L., Rizzo, F., Cremonini, F., Candeloro, E., Luca, L., Grana, C., Bolelli, F.: Do multimodal LLMs understand intraoral dental data? dataset, platform, and baselines. In: European Conference on Computer Vision (2026)
13. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: 2021 IEEE/CVF International Conference on Computer Vision. pp. 82–91 (2021). https://doi.org/10.1109/ICCV48922.2021.00015
14. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. vol. 30, pp. 1195–1204. Curran Associates, Inc. (2017)
15. Wang, Y., Li, Z., Wu, C., Liu, J., Zhang, Y., Ni, J., Luo, Q., Chen, J., Zhang, H., Liu, J., et al.: MICCAI STS 2024 challenge: Semi-supervised instance-level tooth segmentation in panoramic x-ray and CBCT images. Medical Image Analysis p. 103986 (2026)
16. Wang, Y., Zhang, Y., Chen, X., Wang, S., Qian, D., Ye, F., Xu, F., Zhang, H., Dan, R., Zhang, Q., et al.: MICCAI 2023 STS challenge: A retrospective study of semi-supervised approaches for teeth segmentation. Pattern Recognition **170**, 112049 (2026)
17. Wei, L., Sha, R., Shi, Y., Wang, Q., Shi, L., Gao, Y.: Dual consistency semi-supervised learning for 3D medical image segmentation. Biomedical Signal Processing and Control **104**, 107568 (2025). https://doi.org/10.1016/j.bspc.2025.107568
18. Wu, L., Zhuang, J., Chen, H.: VoCo: A simple-yet-effective volume contrastive learning framework for 3D medical image analysis. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22873–22882 (2024). https://doi.org/10.1109/CVPR52733.2024.02158
19. Xu, C., Yang, Y., Xia, Z., Wang, B., Zhang, D., Zhang, Y., Zhao, S.: Dual uncertainty-guided mixing consistency for semi-supervised 3D medical image segmentation. IEEE Transactions on Big Data **9**(4), 1156–1170 (2023). https://doi.org/10.1109/TBDATA.2023.3258643
20. Yang, J., Li, H., Wang, H., Han, M.: 3D medical image segmentation based on semi-supervised learning using deep co-training. Applied Soft Computing **159**, 111641 (2024). https://doi.org/10.1016/j.asoc.2024.111641
21. Yang, X., Ye, M., You, Q., Ma, F.: Writing by memorizing: Hierarchical retrieval-based medical report generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 5000–5009 (2021). https://doi.org/10.18653/v1/2021.acl-long.387