Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Vendor-Aware Three-Dimensional Missing-Tooth Localization in Partially Annotated Multi-Vendor Cone-Beam Computed Tomography: A Leakage-Controlled Evaluation of Semi-Supervision

Hana Asgari
Arak University
Faraneh Zarafshan
Fa.zarafshan@iau.ac.ir

Islamic Azad University, Karaj
Ahmadreza Talaeipoor
Islamic Azad University, Tehran
Elham Mazinan
Arak University

Article

Keywords: cone-beam computed tomography; missing-tooth localization; semi-supervised learning; domain robustness; three-dimensional object detection; dataset integrity

Posted Date: September 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10929754/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Vendor-Aware Three-Dimensional Missing-Tooth Localization in Partially Annotated Multi-Vendor Cone-Beam Computed Tomography: A Leakage-Controlled Evaluation of Semi-Supervision

Hana Asgari*[1], Faraneh Zarafshan†[2,3], Ahmadreza Talaeipoor‡[4], and Elham Mazinan§[1]

[1]Department of Computer, Ar.C., Islamic Azad University, Arak, Iran
[2]Department of Computer Engineering, Ka.C., Islamic Azad University, Karaj, Iran
[3]Institute of Artificial Intelligence and Social and Advanced Technologies, Ka.C., Islamic Azad University, Karaj, Iran
[4]Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran

Abstract

Sparse annotations and scanner-dependent geometry complicate three-dimensional missing-tooth localization in cone-beam computed tomography (CBCT), and unlabeled scans do not necessarily improve detection. We evaluated physical-space reconstruction, vendor-adversarial learning, dense teacher–student consistency, and reliability-gated semi-supervision under leakage-controlled case-level splits. After excluding two discordantly annotated exact duplicates, 83 labeled and 73 unlabeled scans with 111 objects were eligible. Across 75 repeated-seed runs, mean pFROC was 0.890 for pure supervision, 0.899 for vendor-adversarial training, and 0.872 for gated semi-supervision. In 50 post-lock exploratory benchmark runs, conventional 3D detection achieved 0.921 versus 0.886 for confidence-only Mean-Teacher-3D; the primary fold-blocked test was not significant ($p = 0.25$). A cumulative gate ablation showed no pFROC change from uncertainty alone; full gating increased mean pseudo-label precision at confidence 0.70 to 0.826 while reducing coverage to 0.680, and modestly increased pFROC by 0.007 ($p = 0.50$). After fold-specific calibration, vendor-adversarial training achieved the strongest locked outer-test operating point: sensitivity 0.793 (95% CI 0.697–0.880) at 0.458 false positives per volume (0.301–0.639). Post-lock leave-one-vendor-out results were heterogeneous: vendor-adversarial training was numerically strongest for Meyer and Newtom but not Kavo. Semi-supervised superiority was not demonstrated.

Keywords: cone-beam computed tomography; missing-tooth localization; semi-supervised learning; domain robustness; three-dimensional object detection; dataset integrity

1 Introduction

Tooth loss is common and heterogeneous: a scan may contain one or several edentulous sites, and each site must be interpreted relative to neighbouring teeth, alveolar bone, the maxillary sinus, and the mandibular canal. Digital implant workflows increasingly rely on cone-beam computed tomography (CBCT) because it provides volumetric anatomy and enables physical measurements that cannot be recovered reliably from a projection image. Systematic reviews

*Email: hana.asgari@iau.ir
†Correspondence: Fa.zarafshan@iau.ac.ir
‡Email: dr.artalaipour@gmail.com
§Email: Elham.mazinan@iau.ac.ir

nevertheless describe a literature dominated by retrospective, single-centre studies, heterogeneous endpoints, and limited external validation [1, 22, 23, 28]. High reported accuracy in tooth or jaw segmentation does not directly establish accurate localization of an edentulous target, nor robustness across acquisition systems.

Most dental artificial-intelligence pipelines solve adjacent but different tasks. Panoramic-radiograph studies detect, number, or segment visible teeth [3, 5, 15, 40]; CBCT studies predominantly segment teeth, jaws, canals, or sinuses [8, 13, 19, 26, 36, 39, 42]. These advances are important for digital dentistry, yet tooth-instance segmentation presupposes that a tooth is present. A missing-tooth site is instead defined partly by an absence and partly by its three-dimensional anatomical context. Recent implant-planning systems move closer to this objective [11], but planning an implant trajectory is a downstream decision problem and is not equivalent to detecting every candidate missing-tooth region in a multi-site volume.

The released MORPH-CBCT resource directly addresses data scarcity by providing multi-vendor CBCT volumes from patients with missing teeth and three-dimensional boxes for a subset of sites [25]. However, partial annotation creates a nontrivial learning problem. Scans without released boxes cannot be coded as negative: their annotation status indicates unavailable localization labels, not absence of disease or tooth loss. Fully supervised training discards these volumes, whereas naive pseudo-labelling can amplify teacher errors. Semi-supervised learning offers a principled alternative through consistency regularization, exponential-moving-average teachers, confidence filtering, and weak-to-strong perturbations [20, 31, 32, 35, 37, 43]. Medical-imaging studies further show that uncertainty, cross-pseudo supervision, and representation constraints can improve learning with sparse volumetric labels [6, 21, 46–48, 50].

A second difficulty is acquisition shift. CBCT scanners differ in matrix size, voxel spacing, field of view, reconstruction, signedness, rescale metadata, and artifact profile. Deep models may exploit these vendor signatures when label prevalence is imbalanced, creating apparently strong internal results that fail on another scanner. Domain-adversarial learning, covariance alignment, style randomization, and adaptive teacher–student detection provide useful mechanisms [9, 14, 16, 34, 51, 52]; however, removing all vendor information can also suppress clinically useful structure. Robustness must therefore be tested, not assumed, through vendor-stratified evaluation and vendor-held-out sensitivity analysis.

Dataset integrity is equally central. Slice-wise splitting, duplicated volumes, filename-based stacking, and post-hoc exclusions can create leakage or anatomically invalid inputs. Contemporary medical-AI reporting guidance emphasizes explicit data provenance, independent units, prespecified analysis, and transparent subgroup evaluation [7, 18, 24, 29]. We consequently completed a pixel-, series-, geometry-, annotation-, intensity-, and privacy-level audit before model implementation and converted every material finding into a locked decision.

This study asks whether unlabeled CBCT volumes improve three-dimensional missing-tooth localization once data integrity, scanner geometry, and vendor shortcuts are controlled. Rather than assuming that semi-supervision is beneficial, we compare pure supervision, vendor-adversarial learning, dense teacher–student consistency, and reliability-gated consistency under common case-level partitions. The contributions are: (1) a reproducible integrity and eligibility lock that prevents duplicate and slice leakage; (2) geometry-preserving localization in physical coordinates; (3) a controlled evaluation showing that semi-supervised consistency can produce negative transfer even when reliability gating is used; (4) a cumulative gate ablation linking pseudo-label precision and coverage to downstream localization; (5) direct evidence that acquisition-aware vendor control yielded the most favorable observed operating point; and (6) repeated-seed, vendor-stratified, and locked outer-test reporting with case-level uncertainty. The work is therefore positioned as an auditable benchmark and a falsification study of unlabeled-data benefit, not as evidence that the proposed SSL mechanism is superior.

1.1 Related work

This focused narrative review was updated through 14 August 2026 using targeted searches of PubMed-indexed dental and medical-imaging literature and primary computer-vision proceedings. Priority was given to systematic reviews, public CBCT datasets and challenges, externally evaluated dental models, foundational semi-supervised and domain-adaptation methods, and current reporting guidance. It is intended to position the method and define required controls; it is not presented as a PRISMA systematic review.

1.2 Automated dental analysis in two-dimensional radiography

Large panoramic datasets enabled reliable tooth detection, numbering, and segmentation using convolutional detectors and instance-segmentation networks [5, 40, 41]. Missing-tooth classification has also been studied at quadrant level [15]. Projection imaging is computationally convenient and clinically widespread, but superposition and geometric distortion prevent direct recovery of a three-dimensional target box. These studies are therefore informative baselines for dental semantics, not substitutes for volumetric localization.

1.3 CBCT segmentation and three-dimensional dental modelling

Volumetric dental research has progressed from binary tooth/jaw segmentation to individual-tooth identification and multi-structure benchmarks. Hierarchical localization followed by fine segmentation reduces memory cost and class ambiguity [8, 42]; multiclass networks and self-configuring pipelines support teeth, bone, canal, sinus, and restoration-related structures [12, 13, 26, 36, 39]. External and specimen-based validation studies show clinically useful accuracy but also degradation near roots, artifacts, and fine structures [10, 30]. Segmentation is thus a valuable auxiliary representation, but exhaustive voxel labels are more costly than boxes and are not supplied for the unannotated cohort.

1.4 Three-dimensional detection and localization

General object detection evolved from proposal-based detectors to focal-loss and anchor-free formulations [17, 27, 38, 53]. In volumetric medicine, DeepLesion and universal lesion-detection frameworks demonstrated scalable lesion localization from routine CT, while nnDetection automated architecture and training decisions for three-dimensional detection [2, 44, 45]. Dental missing-site localization differs because targets may be small, multiple, and defined by absent anatomy; physical spacing varies markedly between scanners. Accordingly, a detector must operate on geometry-normalized volumes and be evaluated with both overlap and millimetric center error.

1.5 Semi-supervised learning with sparse volumetric labels

Teacher–student learning provides a natural mechanism for exploiting unlabeled scans. Mean Teacher stabilizes targets through exponential weight averaging [37]; FixMatch couples confidence filtering with weak/strong augmentation [31]; detection-specific extensions address region proposals and foreground/background imbalance [20, 32, 35, 43]. Medical SSL adds uncertainty-aware consistency, cross-pseudo supervision, and multi-decoder disagreement to reduce confirmation bias [6, 21, 47, 50]. Yet most medical SSL evidence concerns segmentation. Direct transfer to sparse three-dimensional boxes requires class-aware thresholds, geometric stability checks, and explicit accounting for scans with unknown rather than negative targets.

1.6 Vendor robustness and domain shift

Domain-adversarial neural networks and CORAL align feature distributions [9, 34]; MixStyle perturbs feature statistics to improve domain generalization [52]. Adaptive Teacher, masked

retraining, and class-aware teacher methods combine pseudo-label learning with cross-domain detection [14, 16, 51]. In medical imaging, multi-centre results repeatedly show that acquisition shift can dominate generalization [4, 33, 49]. For the present dataset, vendor and annotation distributions are entangled; consequently, random splitting alone is insufficient. Table 1 distinguishes the evidence streams and the unresolved requirement addressed here.

Table 1: Taxonomy of closely related evidence and the remaining gap.

Evidence stream	Typical input/target	Principal strength	Limitation for this study
Panoramic tooth analysis	2D image; tooth labels/masks	Mature detection and numbering	No physical 3D localization; projection overlap
CBCT segmentation	3D volume; voxel masks	Detailed anatomy and surfaces	Expensive dense labels; assumes structures to segment
Implant planning	CBCT; implant pose/dimensions	Downstream clinical geometry	Often single-site; planning is not exhaustive site detection
General 3D medical detection	CT/MRI; lesion boxes	Scalable volumetric localization	Not designed for absent-tooth context or CBCT vendor shift
Semi-supervised detection	Labeled and unlabeled images	Uses unlabelled data efficiently	Pseudo-box confirmation bias; mostly 2D natural images
Domain adaptation/generalization	Source/target domains	Reduces acquisition shortcuts	Can erase useful anatomy; needs subgroup validation
Present study	Multi-vendor CBCT; partial 3D boxes	Joint localization, SSL, audit, and vendor controls	Repeated-seed and locked outer-fold evidence; no external cohort yet

2 Results

2.1 Dataset integrity and cohort eligibility

The source repository [25] contains three scanner groups (Kavo, Meyer, and Newtom). The audit treated each volume, not each slice, as the independent unit. Table 2 reports verified counts. The 73 scans without released boxes are denoted *unlabeled*; they are not negatives.

Table 2: Verified repository composition before duplicate holdout.

Cohort	Vendor	Scans	DICOM slices	ROI objects
Annotated	Kavo	14	6,032	32
Annotated	Meyer	62	19,840	71
Annotated	Newtom	9	4,550	11
Unlabeled	Kavo	21	9,200	0
Unlabeled	Meyer	32	10,240	0
Unlabeled	Newtom	20	10,115	0
Total		158	59,977	114

Pixel-level fingerprinting found one exact duplicate group: Meyer M38 and M61 shared all 320 slice pixel payloads, physical positions, and SOP instance identifiers, but their released annotations

differed. Both were assigned `hold_out_ambiguous_duplicate` before training. The primary cohort therefore contains 83 labeled scans (Kavo 14, Meyer 60, Newtom 9) and 73 unlabeled scans (Kavo 21, Meyer 32, Newtom 20). Meyer M02 contained a 503-byte text renaming script that is excluded from image manifests. Meyer M73 contains a complete and internally consistent series but filename order is non-monotonic; it remains eligible after geometry-based ordering.

All 114 released annotations are valid oriented boxes in LPS coordinates, lie within their volumes, and represent distinct objects. Multiple files sharing a region code within a scan are spatially distinct objects, not redundant versions. Region counts are R01=16, R02=15, R03=27, and R04=56.

2.2 Prespecified audit domains

The audit was completed before model development and covered archive integrity, file typing, DICOM readability, series identity, physical slice order, spatial geometry, annotation validity, pixel-level duplication, intensity metadata, cohort shift, and privacy risk. Audit findings were translated into explicit inclusion, preprocessing, and reporting decisions (Table 3). This ordering is important: exclusions were not selected after observing model performance.

Table 3: Dataset-integrity findings and their prospective effect on the experimental design.

Audit domain	Verified finding	Locked consequence
Archive and file inventory	59,977 DICOM files, 114 JSON annotations, and one non-DICOM text sidecar	Exclude the sidecar from image manifests; retain the original archive unchanged
Series integrity	158 readable volumes; no duplicate positions, instance numbers, SOP identifiers, instance gaps, or within-volume orientation changes	Require these invariants in every reconstructed volume
Physical ordering	One complete unlabeled series (M73) is not anatomically monotonic in filename order	Sort by projected DICOM position; prohibit filename-based stacking
Pixel duplication	M38 and M61 are exact 320-slice pixel duplicates with discordant released boxes	Hold out both records from primary analysis as group PIX001
ROI geometry	114 boxes are valid, in bounds, and spatially distinct	Treat each JSON file as a separate object; allow multiple objects of one region class
Intensity representation	Vendor-dependent signedness and rescale metadata; no common quantitative scale established	Apply available rescale metadata and training-fitted robust normalization
Privacy	Residual potentially identifying metadata were detected in source DICOM headers	Release only de-identified derived manifests and aggregate audit tables

2.3 Scanner geometry and annotation coverage

Scanner geometry was internally stable but differed materially between vendors (Table 4). Newtom included two acquisition profiles. These differences motivate physical-coordinate localization and vendor-stratified reporting rather than resizing-only preprocessing.

At scan level, 70 labeled scans contained one distinct region class, 13 contained two, and two contained three. Seventeen scans had more than one released ROI object. Pairwise geometry checks found no redundant same-case/same-region boxes (maximum IoU 0.0194) and no overlap among different region codes (minimum center separation 24.6363 mm). Consequently, parenthesized filenames denote multiple spatial objects, not annotation revisions.

Table 4: Verified case-level geometry profiles. Spacing and thickness are in millimetres.

Vendor	Matrix	In-plane spacing	Thickness	Cases	Labeled	Unlabeled
Kavo	640×640	0.250000	0.2500	35	14	21
Meyer	480×480	0.265751	0.2515	94	62	32
Newtom	512×512	0.300000	0.3000	4	included in Newtom totals	
Newtom	820×820	0.150000	0.1500	25	included in Newtom totals	
Newtom total	two profiles	–	–	29	9	20

Table 5: Released three-dimensional ROI objects by vendor and coarse region code.

Vendor	R01	R02	R03	R04	Total
Kavo	10	11	5	6	32
Meyer	6	4	19	42	71
Newtom	0	0	3	8	11
Total	16	15	27	56	114

2.4 Audit results

All 158 volumes passed DICOM reading. No within-volume duplicate positions, duplicate instance numbers, non-contiguous instance sequences, duplicate SOP identifiers, or varying orientations were found. Pixel fingerprinting yielded 157 unique volume fingerprints and isolated only M38/M61 as an exact duplicate pair. All box centers and corners were within volume bounds. Table 6 summarizes the quantitative integrity checks.

Table 6: Quantitative audit summary used to lock preprocessing and eligibility.

Audit measure	Result
Repository cases / DICOM slices	158 / 59,977
Released ROI objects / annotated scans	114 / 85
Eligible labeled / eligible unlabeled scans	83 / 73
DICOM read-error cases	0
Cases with duplicate physical positions / instances / SOP identifiers	0 / 0 / 0
Cases with non-contiguous instance numbers / varying orientations	0 / 0
Unique sequence and slice-multiset pixel fingerprints	157 / 157
Exact pixel-duplicate groups	1 (M38–M61)
ROI centers outside / corners outside one-voxel tolerance	0 / 0
Invalid ROI orientation matrices	0
Same-region pairs with IoU ≥ 0.50 / cross-region pairs with any overlap	0 / 0
Filename-order exceptions	1 (M73)
Non-DICOM sidecars inside image folders	1 (M02)

The slice-spacing audit showed exact 0.2500-mm gaps for Kavo, 0.2514/0.2515-mm quantization for Meyer, and 0.1500- or 0.3000-mm profiles for Newtom. The maximum Meyer deviation from a case median was 0.0001 mm and was interpreted as metadata quantization, not a missing-slice defect. All 114 ROI orientation matrices were identical, orthonormal, and right-handed with determinant 1. Scanner profiles differed in pixel signedness and rescale metadata: Kavo used signed pixels with slope 1 and intercept 0; Meyer used signed pixels without rescale fields; Newtom used unsigned pixels with slope 1 and intercept -1000. Therefore, vendor-aware robustness analysis and training-fitted normalization are mandatory. The developmental validation results

below use the locked eligible cohort and must not be interpreted as outer-test performance.

2.5 Cohort, integrity, and execution

After prospective exclusion of the discordant exact-duplicate pair M38–M61, 156 independent volumes were eligible: 83 labeled scans with 111 retained ROI objects and 73 unlabeled scans. All volumes were reconstructed from DICOM physical geometry and resampled to 0.5-mm isotropic spacing without failure. The five-fold optimization, validation, calibration, and outer-test roles were case-disjoint. All model and evaluation artefacts passed their SHA-256 manifests. Calibration and outer-test partitions were not accessed during model development; after model locking, the outer test was evaluated once.

2.6 Developmental five-fold comparison and ablation

The strict labeled-only detector achieved validation pFROC 0.895 ± 0.084 (mean±SD). The vendor-adversarial reference achieved 0.902 ± 0.088, dense ungated SSL 0.874 ± 0.092, and reliability-gated VR-SSL3D 0.898 ± 0.072 (Table 7). Thus, dense pseudo-labeling reduced mean pFROC by 0.029 relative to the vendor-adversarial reference and degraded four folds (one tie; exact paired Wilcoxon $p = 0.125$). Reliability gating recovered 0.025 pFROC relative to dense SSL, improving three folds, tying one, and degrading one ($p = 0.25$), but was essentially comparable with pure supervision (mean difference +0.003; two improved, one tied, two degraded; $p = 1.00$). These small-sample tests are descriptive and do not establish superiority.

Table 7: Developmental five-fold comparison. pFROC values are mean±SD across inner-validation folds.

Method	Validation pFROC	Difference vs pure	Interpretation
Pure supervised	0.895 ± 0.084	Reference	Strict labeled-only benchmark
Vendor-adversarial	0.902 ± 0.088	+0.007	Small descriptive increase
Dense ungated SSL	0.874 ± 0.092	−0.021	Negative-transfer ablation
Reliability-gated VR-SSL3D	0.898 ± 0.072	+0.003	Recovered dense-SSL loss

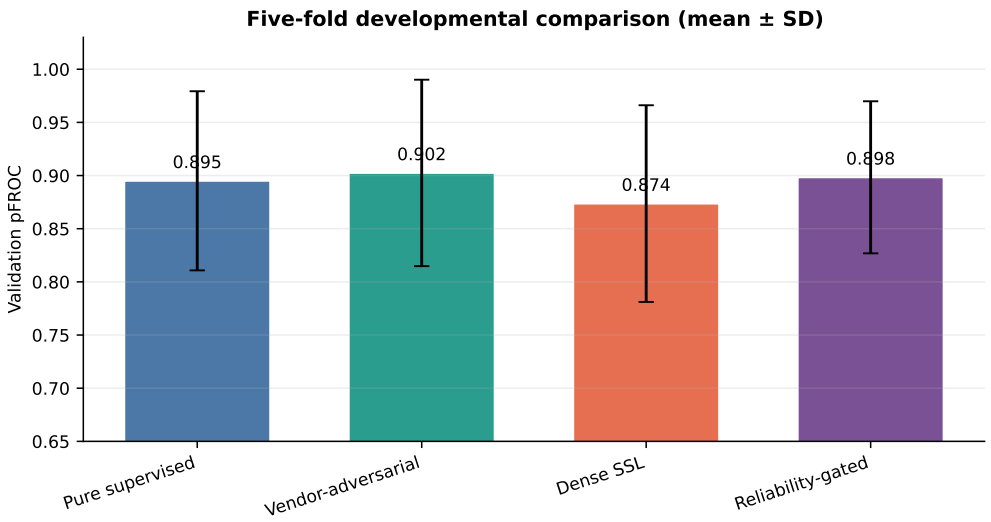


Figure 1: Developmental five-fold pFROC comparison. Bars show fold means and error bars show standard deviations. These values were used for model development and are not outer-test estimates.

2.7 Repeated-seed robustness analysis

To distinguish a stable method effect from a favorable single initialization, we completed five new repeats for each of five folds and three principal regimes (pure supervision, vendor-adversarial training, and reliability-gated VR-SSL3D), yielding 75 of 75 successful runs. No seed was removed or replaced based on performance, and neither calibration nor outer-test data were accessed during this analysis. Dense ungated SSL was not rerun because it served as the completed initial negative-transfer ablation rather than a candidate final model.

Across the 25 fold–repeat outcomes per method, mean pFROC was 0.890 ± 0.087 for pure supervision, 0.899 ± 0.086 for vendor-adversarial training, and 0.872 ± 0.106 for gated VR-SSL3D (Table 8; Fig. 2). Vendor-adversarial minus pure supervision had a mean paired difference of $+0.009$ (11 wins, eight ties, and six losses). Gated VR-SSL3D minus pure supervision had a mean paired difference of -0.019 (two wins, 14 ties, and nine losses), while gated minus vendor-adversarial training was -0.028 (four wins, eight ties, and 13 losses). These paired summaries provide stability evidence but are not patient-level inferential tests: repeat and fold outcomes reuse the same underlying cohort and are therefore not independent clinical samples.

Table 8: Repeated-seed developmental validation. Each method has 25 fold–repeat outcomes (five repeats by five folds). IQR denotes the first to third quartile.

Method	Mean±SD	Median	IQR	Range	n
Pure supervised	0.890 ± 0.087	0.923	0.821–0.954	0.679–1.000	25
Vendor-adversarial	0.899 ± 0.086	0.923	0.804–0.972	0.714–1.000	25
Reliability-gated VR-SSL3D	0.872 ± 0.106	0.923	0.786–0.945	0.625–1.000	25

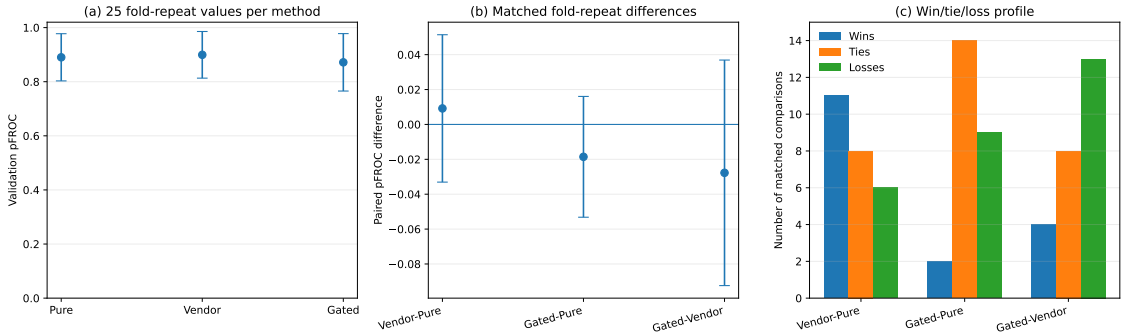


Figure 2: Repeated-seed developmental pFROC distributions from 75 successful runs. The analysis includes five repeats, five folds, and three methods and was completed without calibration or outer-test access. Fold–repeat observations are not independent patient cohorts.

2.8 Post-lock conventional and Mean-Teacher benchmarks

The revision protocol added two architecture-matched exploratory controls after the original outer-test analysis had been locked: a conventional single-scale three-dimensional detector trained only on labeled optimization cases and a confidence-only Mean-Teacher-3D initialized from the corresponding pure-supervised checkpoint. Five repeats across five folds were completed for each control (50 of 50 successful runs); all output manifests passed SHA-256 verification, and calibration and outer-test partitions were not accessed. Conventional 3D achieved mean validation pFROC 0.921 ± 0.065, compared with 0.886 ± 0.078 for Mean-Teacher-3D (Table 9; Fig. 3).

The paired Mean-Teacher-minus-conventional difference was -0.035 (median -0.019; hierarchical bootstrap 95% CI -0.070 to -0.00005), with Mean-Teacher better in three fold–repeat pairs, tied in nine, and worse in 14. Because the five folds, rather than the 25 reused fold–repeat

outcomes, are the defensible independent blocks, the primary exact fold-blocked sign-flip test did not establish a difference ($p = 0.25$). A secondary run-level Wilcoxon test gave $p = 0.003$, but is not treated as primary because repeats nested within folds are dependent. The result therefore supports a descriptive negative-transfer interpretation, not a claim of statistically confirmed superiority for the conventional detector.

Table 9: Post-lock exploratory detector benchmarks. Each method has 25 fold–repeat validation outcomes. These controls cannot alter the locked outer-test ranking.

Method	Mean±SD	Range	n	Time/peak GPU	Unlabeled data
Conventional 3D	0.921 ± 0.065	0.821–1.000	25	11.47 min/1.68 GB	No
Mean-Teacher-3D	0.886 ± 0.078	0.750–1.000	25	11.48 min/1.13 GB	Yes

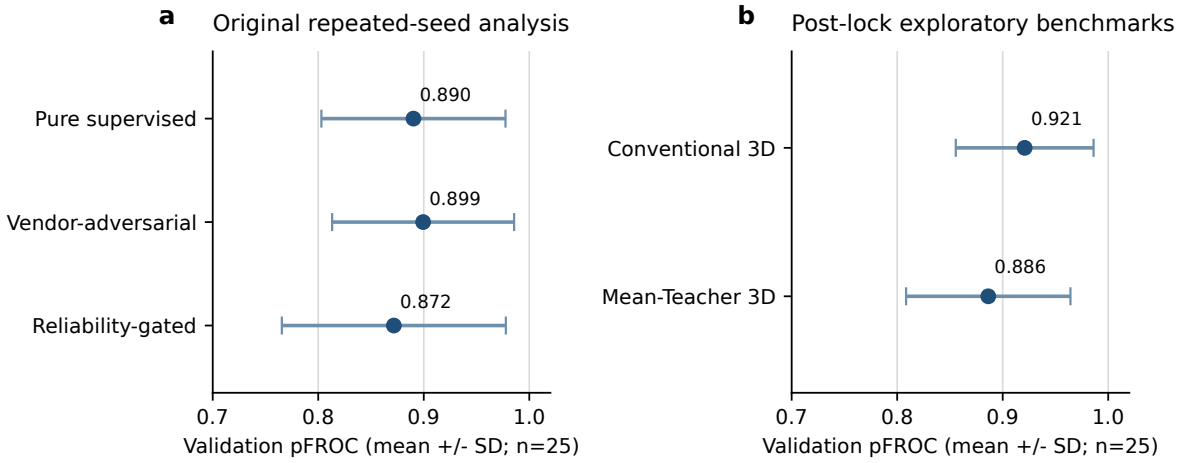


Figure 3: Repeated-seed validation benchmarks. Points show mean pFROC and error bars show standard deviations across 25 fold–repeat outcomes. Panel a reproduces the original three-regime repeated-seed analysis; panel b shows the two post-lock exploratory controls. The panels share folds and repeat counts but belong to different protocol stages and are not presented as a single confirmatory ranking.

2.9 Locked calibration and outer-test performance

For each method and fold, a threshold was selected exclusively in the dedicated calibration partition by maximizing sensitivity subject to no more than 1 false positive (FP) per volume. The locked threshold was then applied unchanged to that fold's outer-test cases. Pooled estimates comprise 83 case-level outer-test evaluations and 111 objects; 95% confidence intervals (CIs) were obtained from 2,000 case-level bootstrap replicates.

The vendor-adversarial model achieved the highest sensitivity, 0.793 (95% CI 0.697–0.880), and the lowest FP burden, 0.458 per volume (95% CI 0.301–0.639), with 88 true positives, 38 false positives, and 23 false negatives (Table 10; Fig. 4). Pure supervision achieved sensitivity 0.757 at 0.602 FP/volume and the smallest median localization error, 1.947 mm. Dense SSL achieved the lowest sensitivity (0.748), consistent with negative transfer observed during validation. Reliability-gated VR-SSL3D achieved sensitivity 0.766 at 0.518 FP/volume: it exceeded dense SSL by 0.018 sensitivity while using a similar FP burden, but it did not outperform the vendor-adversarial model.

Table 10: Locked outer-test performance at fold-specific calibration-selected thresholds. CIs are case-level bootstrap 95% intervals.

Method	Sensitivity (95% CI)	FP/volume (95% CI)	TP/FP/FN	Median error, mm (95% CI)
Pure supervised	0.757 (0.658–0.852)	0.602 (0.422–0.807)	84/50/27	1.947 (1.672–2.174)
Vendor-adversarial	**0.793** (0.697–0.880)	**0.458** (0.301–0.639)	88/38/23	2.163 (1.932–2.350)
Dense ungated SSL	0.748 (0.655–0.838)	0.530 (0.349–0.723)	83/44/28	2.072 (1.836–2.308)
Reliability-gated VR-SSL3D	0.766 (0.672–0.857)	0.518 (0.337–0.735)	85/43/26	2.223 (1.936–2.514)

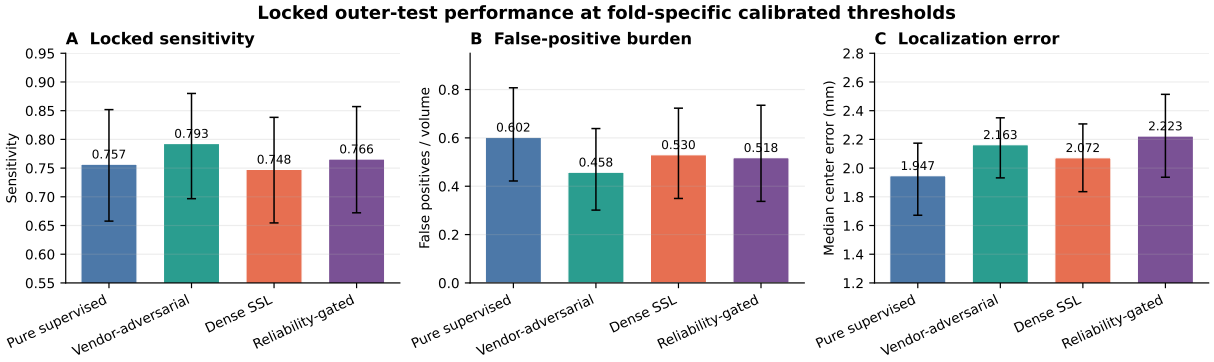


Figure 4: Locked outer-test estimates at fold-specific calibrated thresholds. Error bars denote case-level bootstrap 95% CIs. Higher sensitivity is better; lower false positives per volume and center error are better.

2.10 Post-lock gate-component ablation and pseudo-label quality

To test the mechanism of reliability gating without revisiting calibration or outer-test data, we completed 10 new validation-only training runs: confidence-only and confidence-plus-uncertainty variants in each of five folds. The five previously verified full-gate fold checkpoints were reused, yielding 15 pseudo-label quality evaluations. All outputs passed checksum audit. These analyses are post-lock and exploratory.

Confidence-only and confidence-plus-uncertainty training produced the same mean validation pFROC (0.891 ± 0.077); the fold-level mean difference was effectively zero and the exact five-block sign-flip test gave $p = 1.00$. Adding geometric-stability gating increased mean pFROC to 0.898 ± 0.072, improving folds 0 and 1 and tying folds 2–4 (mean difference +0.007 versus confidence plus uncertainty; exact sign-flip $p = 0.50$). At the common confidence threshold of 0.70, full gating achieved the highest mean pseudo-label precision (0.826), compared with 0.785 for confidence only and 0.723 for confidence plus uncertainty, while accepting a smaller fraction of targets (coverage 0.680 versus 0.710 and 0.696, respectively). Region accuracy among matched pseudo-labels was 1.0 in every fold and regime. Thus, the full gate exhibited the intended precision–coverage trade-off, but neither the localization difference nor the component-specific contribution was statistically established with five fold blocks.

Table 11: Post-lock exploratory cumulative gate ablation across five validation folds. Precision and coverage are measured at the fixed pseudo-label confidence threshold 0.70.

Gate	pFROC, mean±SD	Precision	Coverage	Precision–coverage AUC
Confidence only	0.891 ± 0.077	0.785	0.710	0.525
+ uncertainty	0.891 ± 0.077	0.723	0.696	0.532
+ geometric stability (full)	0.898 ± 0.072	**0.826**	0.680	0.527

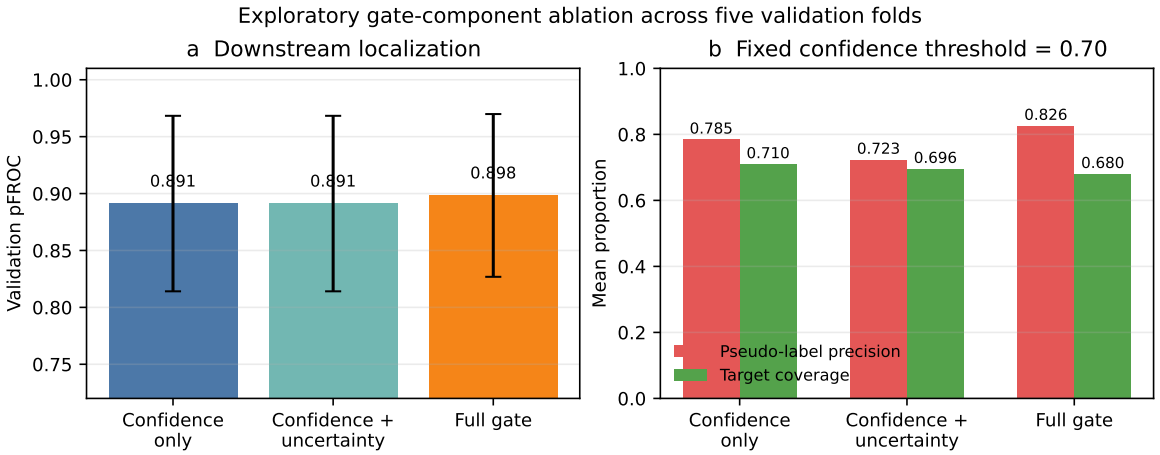


Figure 5: Post-lock exploratory gate-component ablation. (a) Mean validation pFROC with fold-level standard deviations. (b) Mean pseudo-label precision and target coverage at the fixed confidence threshold 0.70. Higher precision under the full gate was accompanied by lower coverage; the five-fold downstream comparison was not significant.

2.11 Vendor-stratified behavior and failure analysis

Vendor-specific estimates were sparse, particularly for Newtom (one to two outer-test cases per fold), and are therefore descriptive. For the gated model, Newtom sensitivity was 1.0 in every fold, but this apparent consistency is based on very small denominators and one detected site had a 4.384-mm center error. Kavo was less stable: fold-specific sensitivity ranged from 0 to 0.80, and the fold-4 Kavo subset contained three cases with zero sensitivity at the locked threshold. Meyer estimates were more stable (fold sensitivity 0.688–1.000) because it contributed 12 cases per fold. These findings identify Kavo fold 4 as a prespecified failure stratum and caution against interpreting pooled performance as scanner-independent robustness.

2.12 Post-lock exploratory strict leave-one-vendor-out evaluation

After the original outer-test analysis had been locked, we completed a strict source-only leave-one-vendor-out (LOVO) domain-generalization experiment requested by the revision protocol. Kavo, Meyer, and Newtom were held out in turn; both labeled and unlabeled scans from the target vendor were excluded from training, model selection, normalization where applicable, and source calibration. Five repeated runs were completed for Pure supervised, vendor-adversarial, and full reliability-gated training, yielding 45 of 45 formal held-out evaluations. All target raw-prediction, provenance, and threshold-count files passed the final SHA-256 evidence audit. No pairwise LOVO superiority test was prespecified, so all comparisons below are descriptive and exploratory.

The held-out cohorts comprised Kavo (14 cases, 32 objects), Meyer (60 cases, 68 objects), and Newtom (9 cases, 11 objects). Across five repeats, mean pFROC for Pure, vendor-adversarial, and reliability-gated training was 0.395, 0.372, and 0.398 when Kavo was held out; 0.786, 0.799, and 0.734 when Meyer was held out; and 0.864, 0.914, and 0.855 when Newtom was held out (Table 12; Fig. 6). The largest favorable pFROC difference for vendor-adversarial training was observed for Newtom (+0.050 versus Pure), followed by Meyer (+0.013), whereas Kavo showed the opposite direction (−0.023 versus Pure). Reliability-gated SSL was essentially tied with Pure for Kavo (+0.003) but lower for Meyer (−0.052) and Newtom (−0.009).

At the source-calibration-selected operating points, vendor-adversarial training showed the highest observed sensitivity for Meyer (0.465) and Newtom (0.818), with mean FP/volume of 0.100 and 0.467, respectively. In Kavo, reliability-gated SSL reduced the mean FP burden to 1.014 FP/volume compared with 1.557 for Pure, but mean sensitivity was lower (0.369 versus

0.394). The source calibration constraint of at most 1 FP/volume applied only when selecting the source-domain threshold; after the threshold was frozen and transferred unchanged, target-domain FP/volume could legitimately exceed 1, as observed for Kavo. Mean center error remained between 1.891 and 2.763 mm across all vendor–method combinations. Bootstrap intervals were broad for the small Kavo and especially Newtom cohorts and are reported in Supplementary Information.

Table 12: Post-lock exploratory strict LOVO performance. Values are mean±SD across five repeated runs. The held-out vendor was excluded from labeled and unlabeled training. No pairwise LOVO superiority test was prespecified.

Held-out vendor	Method	Sensitivity	FP/volume	pFROC	Median center error, mm
Kavo	Pure supervised	0.394 ± 0.075	1.557 ± 0.559	0.395 ± 0.049	2.763 ± 0.352
	Vendor-adversarial	0.381 ± 0.092	1.757 ± 0.577	0.372 ± 0.036	2.612 ± 0.101
	Reliability-gated VR-SSL3D	0.369 ± 0.087	$\mathbf{1.014 \pm 0.523}$	$\mathbf{0.398 \pm 0.043}$	$\mathbf{2.584 \pm 0.153}$
Meyer	Pure supervised	0.385 ± 0.181	$\mathbf{0.090 \pm 0.072}$	0.786 ± 0.046	$\mathbf{2.212 \pm 0.075}$
	Vendor-adversarial	$\mathbf{0.465 \pm 0.186}$	0.100 ± 0.095	$\mathbf{0.799 \pm 0.046}$	2.327 ± 0.065
	Reliability-gated VR-SSL3D	0.412 ± 0.191	0.187 ± 0.298	0.734 ± 0.058	2.227 ± 0.165
Newtom	Pure supervised	0.800 ± 0.119	0.689 ± 0.372	0.864 ± 0.066	$\mathbf{1.891 \pm 0.229}$
	Vendor-adversarial	$\mathbf{0.818 \pm 0.064}$	$\mathbf{0.467 \pm 0.363}$	$\mathbf{0.914 \pm 0.019}$	2.009 ± 0.105
	Reliability-gated VR-SSL3D	0.800 ± 0.076	0.556 ± 0.333	0.855 ± 0.063	1.978 ± 0.088

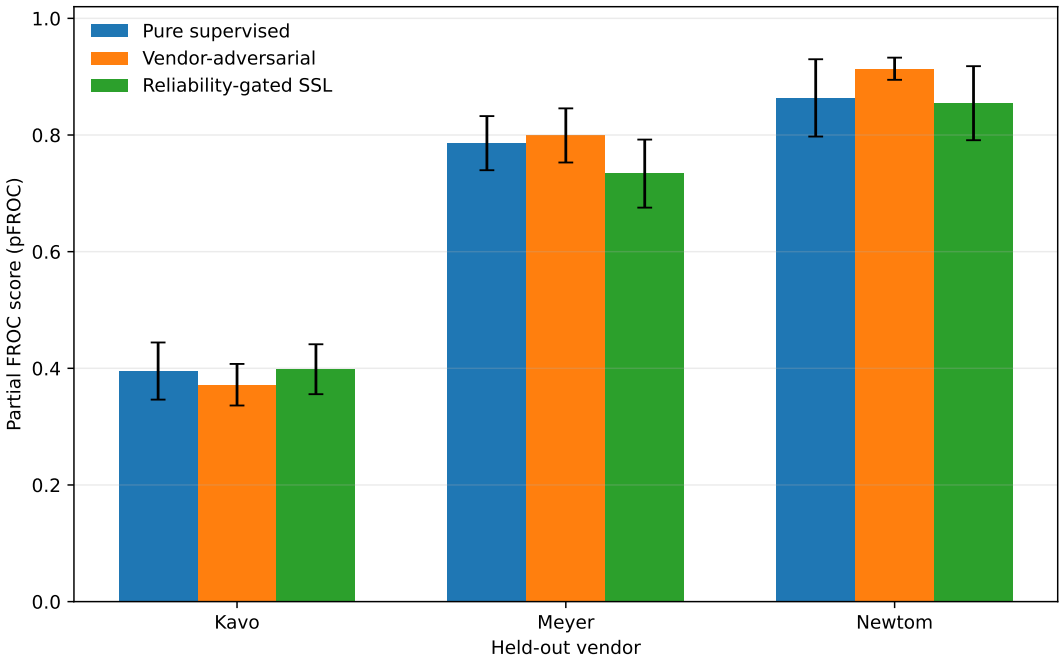


Figure 6: Post-lock exploratory strict LOVO pFROC. Bars show the mean across five repeated runs and error bars show SD. Vendor-adversarial training was numerically most favorable for Meyer and Newtom, whereas Kavo showed no corresponding advantage.

The LOVO findings therefore extend, rather than overturn, the locked analysis. Explicit vendor control can be favorable in some unseen domains, but the effect is not uniform, and reliability-gated use of unlabeled source-domain scans does not provide a consistent held-out-vendor advantage. Because Newtom contributes only nine held-out cases and the Meyer-held-out source calibration partition contains only three cases, these estimates must not be interpreted as proof of universal scanner robustness.

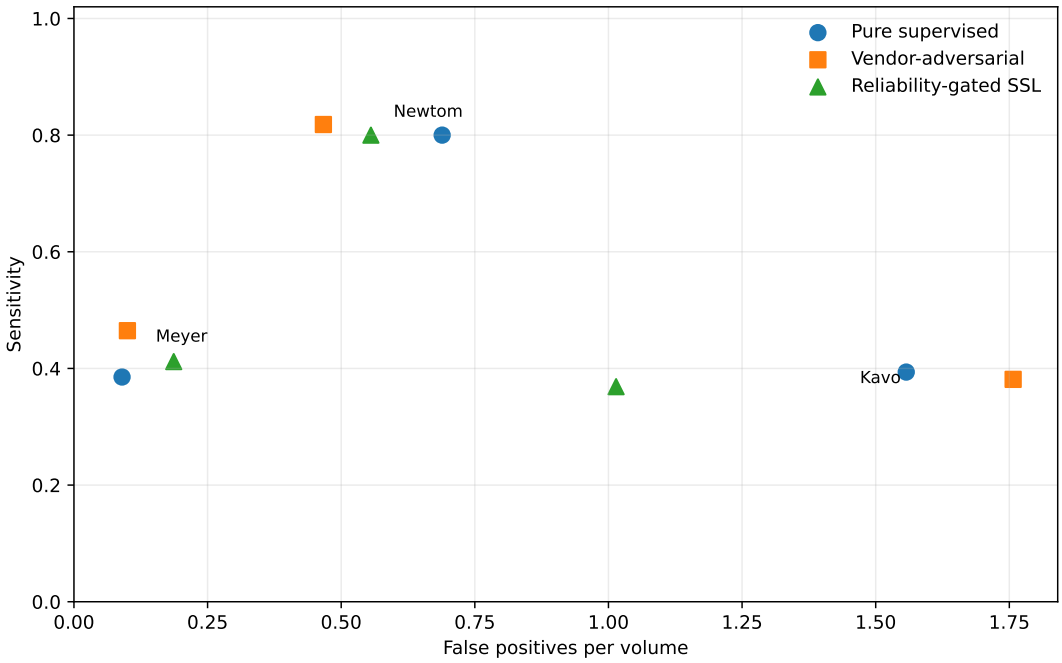


Figure 7: Mean held-out operating points after source-only threshold calibration. Each label denotes the held-out vendor. Vendor-adversarial training showed the most favorable observed sensitivity–FP trade-off for Meyer and Newtom; for Kavo, reliability gating reduced false positives but did not improve sensitivity.

2.13 Qualitative real-CBCT geometry

Figure 8 presents de-identified orthogonal CBCT panels for one annotated case from each scanner family. The overlays are released ground-truth boxes and document scanner-dependent field of view and target geometry; they are not prediction examples. Prediction-level figures remain restricted to the checksum-verified outer-test records and should be added only through the locked selection script, without visual cherry-picking.

2.14 Benchmark and ablation status

The completed evidence now spans the locked developmental comparison, 75 repeated-seed runs, 50 post-lock architecture-matched detector controls, the cumulative gate-component ablation, the locked 83-case outer-test operating point, implementation profiling, and 45 strict LOVO held-out evaluations. The results are directionally coherent in one important respect: none of the semi-supervised variants demonstrates consistent superiority over strong non-SSL controls. Vendor-adversarial training is the most favorable observed method at the original locked outer-test operating point and in the Meyer- and Newtom-held-out LOVO settings, but not in Kavo. Table 13 summarizes the evidential status without collapsing protocol stages into a single confirmatory ranking. Because the post-lock protocol was created after the original outer test had been evaluated, all newly initiated benchmark, gate, and LOVO additions remain exploratory and cannot recast the original comparison as confirmatory evidence.

2.15 Locked operating-point selection and computational profile

For every method and outer fold, the detection threshold was selected using exactly 10 disjoint calibration cases by maximizing sensitivity subject to no more than one false positive per volume. The resulting fold-specific threshold was then applied unchanged to the corresponding outer-test partition. Across folds, the outer-test partitions contained 17, 17, 17, 16, and 16 cases, respectively,

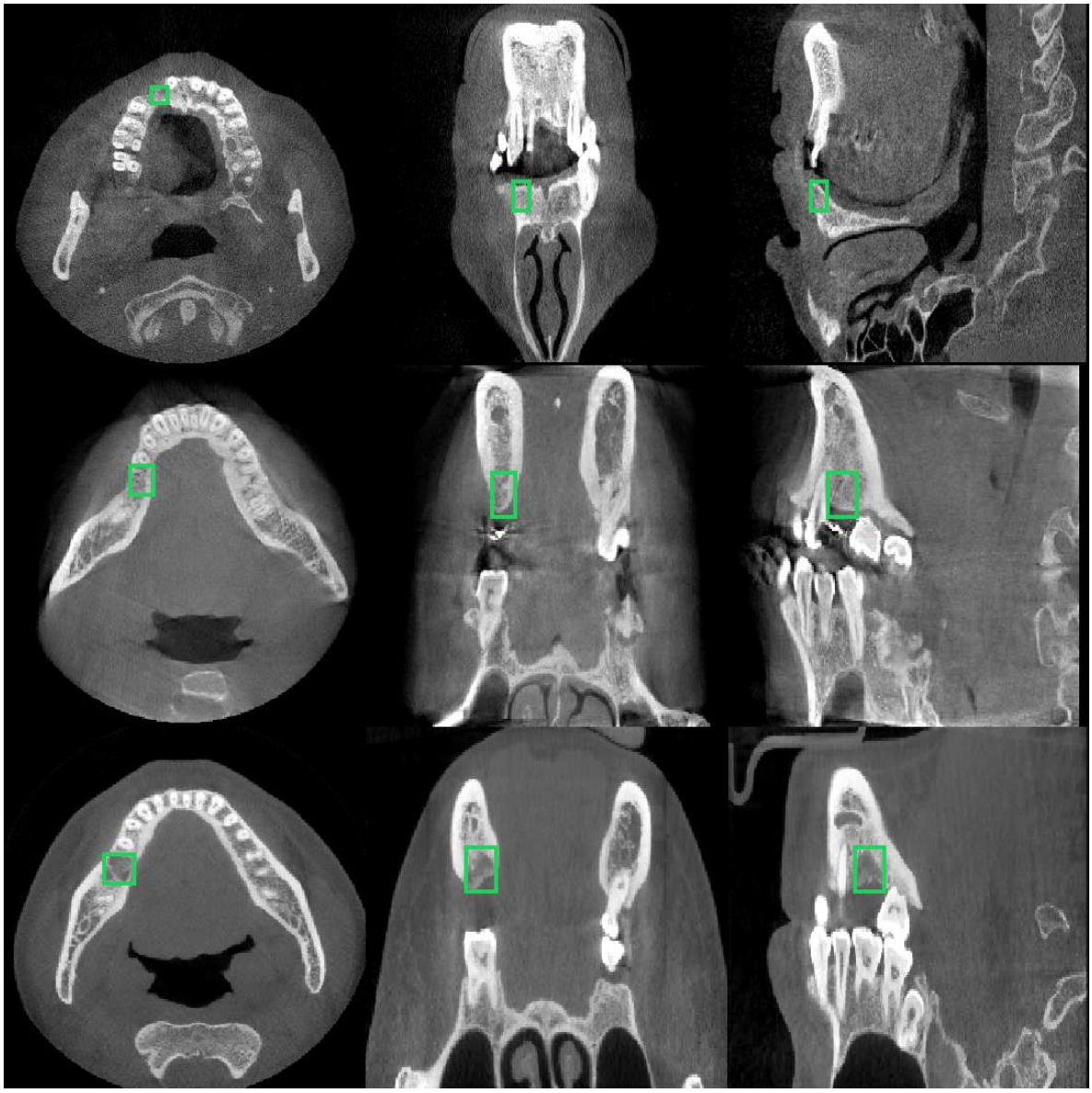


Figure 8: De-identified real CBCT examples from Kavo, Meyer, and Newtom. Columns show orthogonal views and overlays show released ground-truth missing-site boxes.

so every one of the 83 eligible labeled volumes contributed once to pooled out-of-fold evaluation. The outer-test cohort was accessed once after threshold lock. This procedure calibrates the decision operating point; it does not constitute probability calibration and therefore does not support claims about ECE, Brier score, NLL improvement, or calibrated probabilities.

Evaluation of all four original regimes across five folds required 640.5 s (10.7 min) on an NVIDIA GeForce RTX 3080 Ti. Implementation profiling used PyTorch 2.7.0 with CUDA 12.8, mixed-precision FP16, batch size one, gradient checkpointing, and three sequential teacher passes. For the recommended $128 \times 160 \times 160$-voxel patch, peak allocated and reserved memory were 0.723 and 0.781 GB, respectively. Excluding first-call warm-up, measured patch-level execution times for the practical profiled sizes ranged from 0.288 to 0.384 s. These are implementation-level patch measurements, not complete reconstructed-volume or clinical workflow latency.

Table 13: Consolidated benchmark and ablation evidence. Post-lock analyses are exploratory and are not pooled with the original locked outer-test comparison as a single inferential ranking.

Analysis block	Key quantitative result	Defensible interpretation
Developmental five-fold comparison	Pure 0.895; Vendor 0.902; Dense SSL 0.874; Gated 0.898 pFROC	Dense consistency showed negative transfer; gating recovered part of the loss but did not establish superiority.
Repeated-seed robustness (75 runs)	Pure 0.890±0.087; Vendor 0.899±0.086; Gated 0.872 ± 0.106	Vendor was numerically highest; gated SSL did not reproduce a benefit across seeds.
Post-lock detector benchmark (50 runs)	Conventional 0.921 ± 0.065 vs Mean-Teacher 0.886 ± 0.078; fold-blocked $p = 0.25$	Direction supports negative transfer, but conventional-detector superiority was not established.
Gate ablation	pFROC 0.891/0.891/0.898; precision 0.785/0.723/0.826; full-gate $p = 0.50$	Full gating improved pseudo-label precision at reduced coverage without demonstrable downstream superiority.
Locked outer test (83 cases)	Vendor sensitivity 0.793 at 0.458 FP/volume; Pure 0.757/0.602; Dense 0.748/0.530; Gated 0.766/0.518	Vendor-adversarial training provided the most favorable observed locked operating point.
Strict LOVO (45 held-out evaluations)	pFROC Kavo: 0.395/0.372/0.398; Meyer: 0.786/0.799/0.734; Newtom: 0.864/0.914/0.855 (Pure/Vendor/Gated)	Unseen-vendor generalization was heterogeneous; vendor control was favorable for Meyer and Newtom but not Kavo, and SSL showed no consistent benefit.

3 Discussion

The integrity audit materially changes the experimental design. Prospective removal of the ambiguous duplicate pair prevents leakage and avoids post-hoc cherry-picking. Geometry-based sorting prevents anatomically invalid reconstruction of M73. Exclusion of the M02 sidecar prevents a non-image file from entering image counts or loaders. Most importantly, the unlabeled cohort is used only through semi-supervised objectives and is not interpreted as a negative cohort.

The completed locked and exploratory experiments establish seven findings. First, pure supervision was a strong benchmark. Second, dense pseudo-label consistency caused negative transfer in the initial five-fold experiment. Third, the 75-run repeated-seed analysis did not confirm a benefit from reliability gating: gated VR-SSL3D had lower mean pFROC and greater dispersion than pure supervision, although many paired outcomes were ties. Fourth, the post-lock controls yielded the highest validation mean for the conventional detector, whereas confidence-only Mean-Teacher training was lower by 0.035 pFROC on average. This direction was consistent with negative transfer but was not significant in the primary five-block exact test ($p = 0.25$), so it does not establish conventional-detector superiority. Fifth, the cumulative gate ablation showed no downstream change from adding uncertainty alone, while the full geometric-stability gate increased mean pFROC by 0.007 and raised mean pseudo-label precision at confidence 0.70 to 0.826 at the cost of lower coverage; the five-block test was not significant ($p = 0.50$). This supports a precision–coverage mechanism, not superiority. Sixth, vendor-adversarial training showed a small positive repeated-seed mean difference and provided the most favorable locked outer-test operating point, combining the highest sensitivity with the lowest FP burden. The protocol stages, overlapping uncertainty, and small subgroup denominators preclude claims of statistical, scanner-independent, or clinical superiority.

The completed post-lock strict LOVO analysis provides a seventh, complementary finding: generalization was heterogeneous across unseen vendors rather than uniformly improved by either acquisition control or semi-supervision. Vendor-adversarial training had the highest observed mean pFROC when Meyer (0.799) and Newtom (0.914) were held out, but not when Kavo was held out (0.372 versus 0.395 for Pure and 0.398 for gated SSL). In Kavo, the gated model reduced the mean false-positive burden to 1.014 FP/volume versus 1.557 for Pure, but sensitivity was lower (0.369 versus 0.394). Thus, the LOVO evidence supports a vendor-dependent benefit in some domains, not universal scanner robustness, and it provides no consistent evidence that reliability-gated use of unlabeled data improves unseen-vendor localization.

The principal threats are small labeled sample size, vendor–label imbalance, incomplete public

annotations, pseudo-label confirmation bias, and absence of external clinical validation. The 25 fold–repeat observations per method are useful for optimization-stability analysis but are not 25 independent patient cohorts, because folds and repeats reuse the same underlying cases. Accordingly, the fold-blocked test, rather than the nominally stronger run-level result, governs interpretation of the new benchmarks. Vendor-aware performance within this repository does not establish robustness to arbitrary unseen scanners. Localization boxes are coarse spatial targets and do not constitute segmentation masks or treatment recommendations. Conventional-detector, Mean-Teacher, gate-component, and strict LOVO extensions initiated after the locked outer-test analysis are reported only as exploratory. The gate analysis used five validation folds and a fixed 0.70 confidence operating point; its pseudo-label targets are released validation boxes rather than prospectively annotated unlabeled scans. The completed calibration analysis selected fold-specific decision thresholds but did not estimate probability calibration metrics. The computational profile characterizes patch processing rather than full-volume clinical latency. The LOVO analysis is not an independent replication of the original outer test because non-target cases were repartitioned into new source-only development roles; furthermore, the Meyer-held-out source calibration set contained only three cases and the Newtom target contained only nine cases, yielding broad uncertainty. External-cohort analysis remains outstanding. These extensions cannot redefine the primary endpoint or original model ranking.

3.1 Conclusion

This study provides an auditable multi-vendor benchmark for three-dimensional missing-tooth localization and a controlled negative result for semi-supervision. Across 75 original repeated-seed runs, vendor-adversarial training showed a small mean advantage over pure supervision, whereas reliability-gated semi-supervision did not improve mean pFROC. Fifty post-lock exploratory runs further showed a higher descriptive validation mean for conventional 3D detection than for confidence-only Mean-Teacher training, without significance in the primary fold-blocked test. A subsequent cumulative gate ablation found that uncertainty alone did not change downstream pFROC, while the full gate increased pseudo-label precision and modestly recovered localization at reduced coverage without a significant five-fold effect. At locked calibration-selected thresholds, vendor-adversarial training achieved the strongest observed outer-test trade-off (sensitivity 0.793 at 0.458 FP/volume). Post-lock strict LOVO further showed heterogeneous unseen-vendor behavior: vendor-adversarial training was numerically most favorable for Meyer and Newtom but not Kavo, while gated semi-supervision did not show a consistent held-out-vendor benefit. The evidence supports explicit acquisition controls and cautions against assuming that unlabeled localization targets improve performance. It does not establish scanner-independent or clinical superiority.

4 Methods

4.1 Problem definition and design principles

Let $\mathcal{D}_L = \{(x_i, Y_i, v_i)\}_{i=1}^{N_L}$ denote the eligible labeled CBCT volumes and $\mathcal{D}_U = \{(u_k, v_k)\}_{k=1}^{N_U}$ the volumes for which no three-dimensional boxes were released. Here, $v \in \{\text{Kavo}, \text{Meyer}, \text{Newtom}\}$ is the scanner vendor and

$$Y_i = \{(b_{ij}, r_{ij})\}_{j=1}^{m_i}, \qquad b_{ij} = (x, y, z, w, h, d)_{ij} \in \mathbb{R}^6, quadr_{ij} \in \{\text{R01}, \ldots, \text{R04}\}, \tag{1}$$

where (x, y, z) and (w, h, d) are the box center and extent in millimetres. The absence of a released box for u_k is treated as missing supervision, not evidence of a negative volume. The objective is to learn a detector f_θ that returns a set of scored, class-labelled three-dimensional missing-site boxes while minimizing sensitivity to vendor-specific acquisition signatures.

The proposed VR-SSL3Dframework follows four locked principles: (i) all geometric operations are defined in physical patient space; (ii) labeled supervision acts on the student detector, whereas

unlabeled targets are produced by a no-gradient exponential-moving-average (EMA) teacher; (iii) a pseudo-box is accepted only when confidence, predictive uncertainty, and transformation stability all pass prespecified gates; and (iv) vendor invariance and probability calibration are separated, with calibration fitted only on a held-out calibration partition and never on the final test set.

4.2 Geometry-preserving CBCT reconstruction

For slice i, the DICOM row and column direction cosines $\mathbf{r}_i$ and $\mathbf{c}_i$ define the slice normal $\mathbf{n}_i = \mathbf{r}_i \times \mathbf{c}_i$. With $\mathbf{p}_i$ denoting `ImagePositionPatient`, slices are ordered by

$$s_i = \mathbf{p}_i^\top \mathbf{n}_i, \tag{2}$$

not by filename. Duplicate positions, inconsistent orientations, and non-unique SOP identifiers are fatal reconstruction errors. Available rescale slope and intercept are applied, but scanner values are not assumed to be quantitative Hounsfield units. Robust clipping parameters are estimated from the training partition only, followed by per-volume normalization.

The DICOM affine $A_i \in \mathbb{R}^{4\times 4}$ maps a homogeneous voxel coordinate $\tilde{\boldsymbol{\xi}} = (c, r, s, 1)^\top$ to patient space:

$$\tilde{\mathbf{x}}_{\mathrm{mm}} = A_i \tilde{\boldsymbol{\xi}}, \qquad A_i = \begin{bmatrix} \Delta_c \mathbf{c}_i & \Delta_r \mathbf{r}_i & \Delta_s \mathbf{n}_i & \mathbf{p}_0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

Every volume is resampled to a prespecified isotropic spacing using trilinear interpolation for intensities; box corners are transformed analytically through the same affine. Resampling therefore changes the computational lattice but not the physical target. The network input has shape $B \times 1 \times D \times H \times W$, and predicted boxes are converted back to millimetres before non-maximum suppression or evaluation.

4.3 Student detector and EMA teacher

As summarized in Fig. 9, the student contains a three-dimensional encoder, a feature-pyramid network (FPN) producing $\{P_2, P_3, P_4, P_5\}$, and anchor-free heads for objectness, box regression, and region classification. The three heads are shared across pyramid levels. A candidate a has objectness logit $z_{o,a}$, box $\hat{b}_a$, and region logits $\mathbf{z}_{r,a} \in \mathbb{R}^4$. Multi-scale prediction is required because a site extent in voxels changes with both physical size and resampling.

The teacher has the same architecture and is not optimized by gradient descent. After student update t, its parameters are

$$\bar{\theta}_t = m\bar{\theta}_{t-1} + (1-m)\theta_t, \tag{4}$$

where $m \in [0, 1)$ is the EMA momentum. The student is used for final inference; the teacher is used only to construct stable unlabeled targets and may be reported as a prespecified sensitivity model.

4.4 Supervised localization objective

Ground-truth boxes are assigned to FPN locations using a fixed center-sampling rule in physical space. Let $y_{o,a} \in \{0, 1\}$ be the objectness target and $\mathcal{P}$ the matched positive locations. To reduce foreground–background imbalance, the objectness component uses focal binary cross-entropy:

$$\mathcal{L}_{obj} = \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} -\alpha_{y_a}(1 - p_{t,a})^\gamma \log p_{t,a}. \tag{5}$$

Box regression combines three-dimensional generalized IoU and normalized center distance,

$$\mathcal{L}_{box} = \frac{1}{|\mathcal{P}|} \sum_{a\in\mathcal{P}} \left[1 - \mathrm{GIoU}_{3D}(\hat{b}_a, b_a) + \eta \frac{\|\hat{\mathbf{c}}_a - \mathbf{c}_a\|_2}{\sqrt{w_a^2 + h_a^2 + d_a^2} + \epsilon} \right], \tag{6}$$

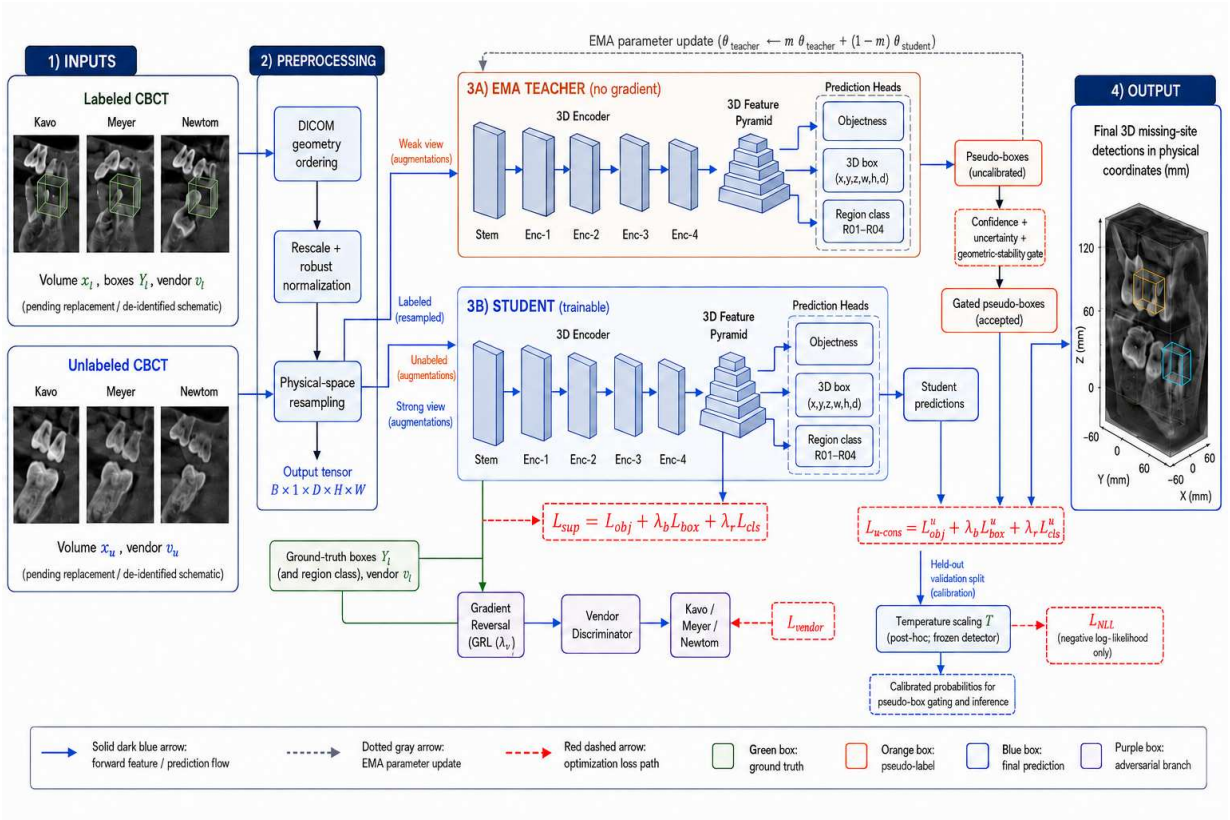


Figure 9: Overview of VR-SSL3D. Labeled volumes supervise the trainable student; weak and strong views of unlabeled volumes are processed by the EMA teacher and student, respectively. Solid blue arrows denote forward flow, dotted grey arrows the EMA update, and red dashed arrows optimization paths. The three losses attached to the student heads form $\mathcal{L}_{sup}$; accepted teacher pseudo-boxes and student predictions form $\mathcal{L}_{u\text{-}cons}$; and multi-scale student features feed the gradient-reversal vendor branch. Figure 10 expands pseudo-box gating. Fold-specific decision thresholds are selected only after training on disjoint calibration cases.

and $\mathcal{L}_{cls}$ is class-weighted cross-entropy over R01–R04. Class weights are computed from the training fold only. The supervised objective is

$$\mathcal{L}_{sup} = \mathcal{L}_{obj} + \lambda_b \mathcal{L}_{box} + \lambda_r \mathcal{L}_{cls}. \tag{7}$$

4.5 Reliability-gated three-dimensional pseudo-boxes

Figure 10 details the principal semi-supervised contribution. For an unlabeled volume u, the teacher receives K weak stochastic views $T_w^{(k)}(u)$. Predictions are inverse-transformed to the same physical frame and associated by class and three-dimensional IoU. For candidate a of predicted class r, the joint confidence is

$$q_a = \bar{q}_{o,a}\,\bar{p}_{r,a}, \tag{8}$$

where bars denote the mean across matched teacher predictions. Predictive uncertainty combines normalized center, log-extent, and class-probability dispersion:

$$u_a = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{\|\mathbf{c}_a^{(k)} - \bar{\mathbf{c}}_a\|_2^2}{s_0^2} + \rho_b\|\log \mathbf{e}_a^{(k)} - \log \bar{\mathbf{e}}_a\|_2^2 + \rho_p\|\mathbf{p}_a^{(k)} - \bar{\mathbf{p}}_a\|_2^2\right], \tag{9}$$

with $\mathbf{e} = (w, h, d)$ and normalization scale s_0 fixed before testing.

Geometric stability is evaluated using a paired augmentation with known transform T. A box predicted in view B is mapped back to view A, $b_a^{B\to A} = T^{-1}(b_a^B)$. The stability measures are

$$d_{mm,a} = \|\mathbf{c}_a^A - \mathbf{c}_a^{B\to A}\|_2, \qquad j_a = \mathrm{IoU}_{3D}(b_a^A, b_a^{B\to A}). \tag{10}$$

Candidate a is accepted exactly when

$$g_a = \mathbb{1}\Big[q_a \geq \tau_c^{(r)}\Big]\,\mathbb{1}\Big[u_a \leq \tau_u^{(r)}\Big]\,\mathbb{1}\Big[d_{mm,a} \leq \tau_d^{(r)}\Big]\,\mathbb{1}\Big[j_a \geq \tau_{IoU}^{(r)}\Big] = 1. \tag{11}$$

Thresholds are tuned on validation data with a minimum pseudo-box precision constraint; they are never chosen from test outcomes. Accepted boxes undergo class-wise NMS in millimetres and are transformed into the strong student view $T_s(u)$. Rejected candidates are ignored and never converted into negative examples.

Let $\widehat{Y}_u$ denote accepted teacher targets and Y_s the matched student predictions. The consistency objective mirrors Eq. 7:

$$\mathcal{L}_{u\text{-}cons} = \mathcal{L}_{obj}^u + \lambda_b \mathcal{L}_{box}^u + \lambda_r \mathcal{L}_{cls}^u. \tag{12}$$

The denominator is the number of accepted pseudo-boxes (with a safe zero-target convention). This prevents volumes with many low-quality proposals from dominating an update.

4.6 Geometry-consistent weak and strong augmentation

The weak transform T_w is limited to intensity perturbations and small rigid transformations that preserve dental anatomy. The strong transform T_s additionally permits stronger gamma/contrast perturbation, Gaussian noise or blur, bounded cutout outside matched target centers, and larger but anatomically plausible rigid transformations. Any spatial transform is represented by an invertible 4×4 affine and is applied identically to the volume and boxes. Left–right reflection is disabled because it can change the semantic relation between region labels unless a validated label-remapping rule is supplied. Elastic deformation, anisotropic scaling, and cropping that truncates an accepted box are prohibited in the primary experiment.

All augmentation ranges are fixed using training and calibration data. To prevent the strong view from becoming an artificial domain, the realized distribution of rotations, translations, retained field of view, and intensity percentiles is logged by vendor. A pseudo-box is excluded from the unsupervised loss if less than a prespecified fraction of its physical volume remains after transformation.

4.7 Candidate association and physical-space suppression

Teacher predictions from stochastic views are associated only when their region class agrees and their inverse-transformed boxes exceed a matching threshold τ_{match}. The association is solved greedily in descending joint confidence, with each candidate used at most once. The representative pseudo-box is the confidence-weighted mean of centers and log-extents:

$$\bar{\mathbf{c}}_a = \frac{\sum_k q_a^{(k)} \mathbf{c}_a^{(k)}}{\sum_k q_a^{(k)}}, \qquad \log \bar{\mathbf{e}}_a = \frac{\sum_k q_a^{(k)} \log \mathbf{e}_a^{(k)}}{\sum_k q_a^{(k)}}. \tag{13}$$

Class-wise NMS is performed after conversion to patient-space millimetres. This ordering avoids scanner-dependent decisions caused by different native voxel spacings. During supervised assignment, every released ROI remains a distinct object; repeated region codes within one scan are never merged solely because their class labels match.

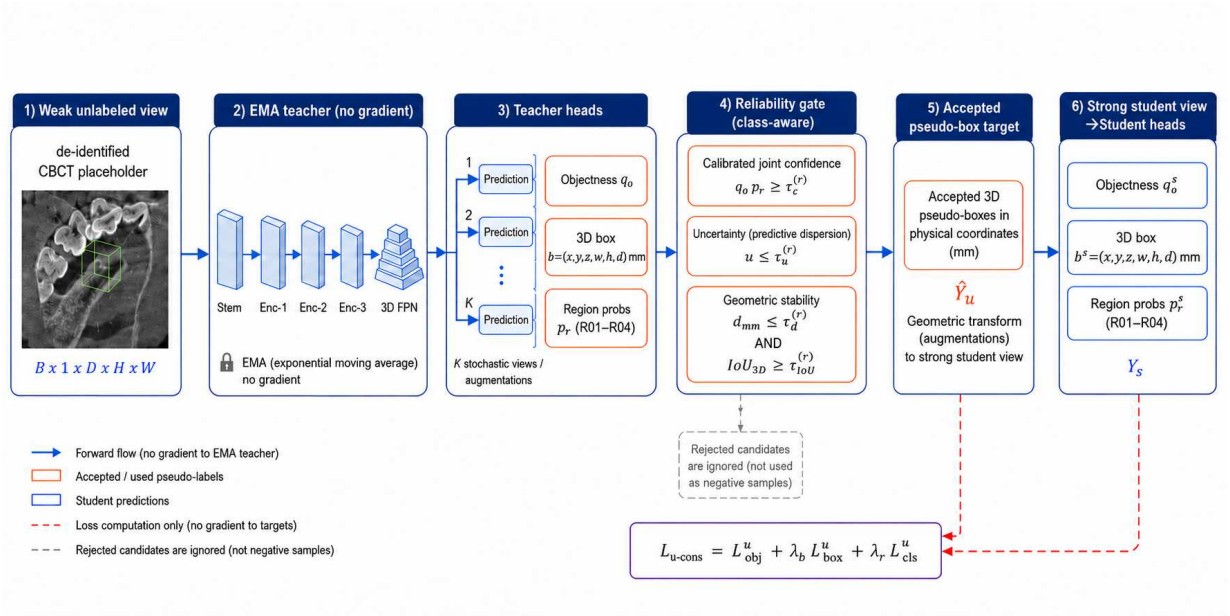


Figure 10: Reliability-gated pseudo-box generation. K stochastic weak views yield teacher confidence, predictive dispersion, and transformation-stability estimates. A candidate must pass all class-aware gates in Eq. 11; accepted boxes are transformed into the strong student frame and supervise Eq. 12. Rejected candidates are ignored rather than used as negatives. All box operations are performed in millimetres.

4.8 Vendor-invariant representation learning

The vendor-adversarial branch discourages scanner shortcuts. Each pyramid tensor is spatially pooled, the four vectors are concatenated and normalized to produce z_v. A discriminator d_ϕ predicts the vendor after a gradient-reversal layer GRL_{λ_v}:

$$\hat{v} = d_\phi(\text{GRL}_{\lambda_v}(z_v)), \qquad \mathcal{L}_{vendor} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{3} w_c^v \mathbb{1}[v_i = c] \log \hat{v}_{ic}. \tag{14}$$

The discriminator minimizes this cross-entropy, whereas reversal multiplies the gradient entering the encoder by $-\lambda_v$. Vendor-balanced minibatches and weights w_c^v prevent the majority vendor from defining the adversarial signal. The branch is applied to labeled and unlabeled training volumes because vendor labels are available for both.

4.9 Leakage-safe operating-point calibration

Within each fold, detector scores on the disjoint calibration partition were swept over candidate thresholds. The locked threshold τ_f^* was selected by

$$\tau_f^* = \arg\max_{\tau} \text{Sensitivity}_f(\tau) \quad \text{subject to} \quad \text{FP/volume}_f(\tau) \leq 1. \tag{15}$$

Each calibration partition contained 10 cases and was excluded from detector-gradient updates. The selected threshold was applied unchanged to the corresponding outer-test fold, which was accessed once after all fold-specific thresholds had been locked. This is decision operating-point calibration rather than probability calibration: no temperature parameter was fitted, and ECE, Brier score, NLL improvement, and probability-reliability claims are not reported.

4.10 Complete objective and optimization schedule

The detector-stage objective is

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_u(t)\mathcal{L}_{u\text{-}cons} + \lambda_v(t)\mathcal{L}_{vendor}, \tag{16}$$

where $\lambda_u(t)$ is zero during supervised warm-up and then follows a sigmoid ramp-up; $\lambda_v(t)$ is likewise warmed up to avoid erasing useful features before localization is established. Operating-point selection is intentionally absent from Eq. 16 because it is a post-training evaluation procedure and does not update detector parameters.

Table 14: Algorithm 1: leakage-controlled optimization of VR-SSL3D.

1	Lock case eligibility, duplicate groups, and vendor-stratified splits.
2	Fit preprocessing statistics on training cases and reconstruct every series by DICOM geometry.
3	Warm up the student on $\mathcal{D}_L$ using $\mathcal{L}_{sup}$; update the teacher by EMA.
4	Keep the disjoint calibration and outer-test partitions inaccessible during detector fitting and model selection.
5	Sample vendor-balanced labeled and unlabeled minibatches; create weak teacher and strong student views.
6	Compute $\mathcal{L}_{sup}$ on labeled student predictions.
7	Generate K teacher predictions, inverse-transform them, compute q, u, d_{mm} and IoU_{3D}, and retain only candidates satisfying Eq. 11.
8	Transform accepted boxes to the student frame and compute $\mathcal{L}_{u\text{-}cons}$; ignored candidates produce no negative target.
9	Compute $\mathcal{L}_{vendor}$ from both cohorts and update the student/discriminator with Eq. 16.
10	Update the EMA teacher and evaluate checkpoint-selection criteria only on validation data.
11	Select one checkpoint, choose its threshold on the disjoint calibration partition under the locked FP constraint, and evaluate the outer-test fold once.

4.11 Inference and post-processing

At inference, a reconstructed volume is processed once by the selected student. Decoded boxes are mapped to patient-space millimetres, and class-wise three-dimensional NMS removes redundant predictions. A detection is retained when its score exceeds the fold-specific calibration-selected threshold. No vendor label is required by the localization path. Outputs comprise case identifier, vendor for stratified analysis, region class, detector score, center, extent, and processing time. The system localizes candidate missing sites; it does not segment bone, prescribe implant dimensions, or make a treatment decision.

4.12 Implementation contract and reproducibility

Before full training, the implementation must pass deterministic unit tests for DICOM slice ordering, affine round trips, box transformation, three-dimensional IoU, center distance, matching, NMS, and rejection of truncated boxes. A synthetic cuboid transformed from voxel to patient coordinates and back must recover every corner within a fixed numerical tolerance. M73 is a mandatory regression test for geometry-based ordering, whereas M38 and M61 must be absent from every optimization, calibration, and primary test manifest.

For each run, the exported configuration records input spacing and crop policy; backbone and FPN widths; optimizer, learning-rate schedule, batch composition, mixed precision, gradient

accumulation, and clipping; EMA momentum; all loss weights; K; every gate and NMS threshold; calibration protocol; random seed; software versions; hardware; and the checksums of split manifests and source code. Case-level predictions are saved before metric computation so all tables can be regenerated without rerunning inference.

4.13 Computational complexity

Let $V = DHW$ be the resampled voxel count and C_l the channel width of layer l. A single student pass has the usual convolutional cost $\mathcal{O}(\sum_l V_l k_l^3 C_{l-1} C_l)$. Supervised training requires one student pass per labeled view. Each unlabeled update uses one strong student pass and K no-gradient teacher passes; therefore the principal SSL time multiplier is approximately $1+K$ for the unlabeled part, while activation memory is dominated by the student because teacher activations are not retained for backpropagation. The vendor discriminator adds only pooled vectors and two fully connected layers, and its cost is negligible relative to the 3D encoder. Profiling was performed on an NVIDIA GeForce RTX 3080 Ti (11.63 GB), using PyTorch 2.7.0, CUDA 12.8, FP16, batch size one, gradient checkpointing, and sequential teacher passes. The recommended $128 \times 160 \times 160$ patch reserved 0.781 GB. Patch-level steady-state timings are reported separately from end-to-end volume latency.

4.14 Failure safeguards and falsifiable ablations

Confirmation bias is monitored through pseudo-box acceptance rate, class/vendor distribution, validation precision, center disagreement, and uncertainty over training. If no candidate passes the gate, $\mathcal{L}_{u\text{-}cons} = 0$ for that minibatch; thresholds are not relaxed online using test information. A minimum supervised proportion is retained in every update. Vendor-adversarial collapse is monitored by localization validation performance together with discriminator accuracy: chance-level vendor prediction is not considered success if anatomical localization degrades.

Each claimed contribution has a prespecified falsification experiment. Geometry awareness is tested against voxel-only resizing; the reliability gate against confidence-only, confidence-plus-uncertainty, and full gating; semi-supervision against the identical labeled-only detector; vendor robustness against removal of GRL and vendor balancing; physical-space regression against voxel-space box regression; and calibration against uncalibrated scores. The method will not be described as vendor-robust unless gains are consistent in vendor-stratified results and the vendor-held-out sensitivity analysis, with uncertainty intervals reported.

4.15 Experimental protocol

4.16 Prespecified estimand and evaluation unit

The primary estimand is the difference in case-level three-dimensional localization performance between VR-SSL3Dand the architecture-identical supervised detector when both are trained and selected under the same labeled-data budget. The independent sampling unit is the CBCT volume. Slices, boxes, and augmented views from one volume can never occur in different partitions. M38 and M61 remain excluded as the locked ambiguous duplicate group. The 73 unannotated volumes are training resources only and are never counted as negative cases or used to calculate localization metrics.

Because only 83 eligible labeled volumes remain and the smallest vendor group contains nine labeled scans, the primary analysis uses locked five-fold outer cross-validation rather than a single small holdout. Folds are generated once by constrained case-level stratification over vendor, number of ROI objects, and presence of R01–R04, subject to feasibility. The optimizer must report the achieved fold balance and any unsatisfied constraint; it may not silently move cases after performance inspection. Each labeled case is an outer-test case exactly once. All model design, thresholds, and stopping decisions for an outer fold use only its complementary development cases.

4.17 Nested development, calibration, and unlabeled-data use

Within each outer development set, a fixed inner split separates optimization, validation, and operating-point calibration roles. Class weights, augmentation ranges, and detector parameters are fitted on optimization cases. Validation cases determine early stopping, pseudo-box gates, NMS, and checkpoint selection. The 10 calibration cases in each fold are used only to select the decision threshold under the locked constraint of at most one false positive per volume. The outer-test fold remains untouched until that threshold has been fixed.

All 73 eligible unlabeled volumes may be used in the optimization portion of every outer fold because they are distinct cases with no released localization endpoint. Their vendor identities may supervise $\mathcal{L}_{vendor}$, but their absent box files provide no negative target. Normalization followed the executed method-specific protocol: Pure supervised runs fitted statistics from labeled optimization cases only; vendor-adversarial runs fitted statistics from labeled optimization cases plus eligible unlabeled cases; and reliability-gated runs used the byte-identical normalization statistics of the corresponding Pure run while still using eligible unlabeled scans for vendor and consistency objectives. The same method-specific normalization contract was preserved in the repeated-seed and strict LOVO extensions. No normalization statistic used a held-out target vendor or an outer-test labeled volume. A supervised baseline receives the same labeled optimization cases, augmentations, number of update opportunities, checkpoint rule, and backbone, but no unlabeled image gradient.

4.18 Original comparisons and post-lock extensions

The completed, locked analysis compared pure supervision, vendor-adversarial supervision, dense teacher–student consistency, and reliability-gated VR-SSL3D. The original hypothesis that VR-SSL3Dwould improve upon the identical supervised detector was not supported. It is therefore reported as a falsified developmental hypothesis rather than retained as a confirmatory superiority claim. The following post-lock extensions address mechanism and benchmark completeness:

1. **nnDetection/standard 3D detector**: a reproducible conventional volumetric detector configured for the same resampled inputs and folds;
2. **Mean-Teacher-3D**: EMA teacher with confidence-filtered pseudo-boxes, without uncertainty, geometric stability, vendor adversarial learning, or calibration;
3. **Reliable-SSL3D**: full reliability gate but no vendor discriminator;
4. **Vendor-Adv-3D**: supervised detector with GRL vendor branch but no unlabeled localization consistency;
5. **VR-SSL3D**: complete proposed method.

All trainable extensions use the same immutable case manifests, physical-space preprocessing, evaluation code, fixed seed set, and approximately matched optimization budget. A baseline is not disadvantaged by using a weaker backbone unless that architecture is intrinsic to the published method, in which case original and backbone-matched variants are distinguished. Because these extensions were specified after the original outer-test results were known, development and calibration partitions may be used under the revision lock, but no hyperparameter, checkpoint rule, or gate can be chosen using outer-test outcomes. Any subsequent application to the previously accessed outer-test cohort is labelled exploratory and cannot replace the original locked comparison.

4.19 Ablation and mechanism tests

The completed cumulative gate sequence was architecture-matched: (A0) supervised detector; (A1) EMA plus confidence gating; (A2) A1 plus predictive-uncertainty gating; and (A3) A2

Table 15: Locked comparison matrix for the primary experiments.

Method	Unlabeled CBCT	EMA teacher	Full gate	Vendor GRL	Calibration
Supervised-3D	–	–	–	–	optional report-only
Mean-Teacher-3D	yes	yes	confidence only	–	–
Reliable-SSL3D	yes	yes	yes	–	yes
Vendor-Adv-3D	vendor labels only	–	–	yes	yes
VR-SSL3D	yes	yes	yes	yes	yes

plus physical geometric-stability gating. Vendor-balanced sampling and GRL were evaluated in the separate vendor-adversarial regime. Each completed gate comparison reports downstream localization together with pseudo-box precision, coverage, center error, and per-vendor behavior so an apparent gain cannot be attributed merely to accepting more pseudo-labels.

Table 16: Claims, falsification tests, and required evidence.

Claim	Direct test	Evidence required
Unlabeled data improve localization	VR-SSL3D versus identical supervised detector	Paired outer-fold/seed difference with interval; no increase in FP/volume
Reliability gating controls confirmation bias	Confidence-only versus uncertainty and full gates	Pseudo-box precision, coverage, center error, and downstream localization
Physical coordinates reduce scanner dependence	Millimetric versus voxel-space box regression/NMS	Vendor-stratified center error and IoU_{3D}
Vendor branch reduces shortcuts	With versus without balanced sampling and GRL	Localization retained/improved while vendor predictability decreases
Operating-point selection controls FP burden	Calibration-selected versus arbitrary score threshold	Sensitivity and FP/volume after unchanged application to the outer fold

4.20 Detection matching and primary endpoint

Predictions are matched one-to-one to released boxes within the same volume and region class. Candidate pairs are ranked by three-dimensional IoU and assigned greedily (a Hungarian assignment is a prespecified sensitivity analysis). A true positive must satisfy $IoU_{3D} \geq \delta_{IoU}$; because small boxes can incur a large IoU penalty for a clinically modest displacement, a second prespecified analysis uses center distance $d_{mm} \leq \delta_d$. Thresholds δ_{IoU} and δ_d are chosen from clinical/annotation considerations before outer-test evaluation and are not optimized per method.

The primary endpoint is free-response localization sensitivity at a fixed false-positive burden per volume, selected before analysis. The complete FROC curve and partial area over the prespecified FP/volume range are also reported. This endpoint prioritizes finding missing sites without concealing the cost of multiple false detections. Sensitivity is computed from pooled out-of-fold predictions, with the case as the resampling unit.

4.21 Secondary endpoints

Secondary localization endpoints are AP_{3D} at the locked IoU threshold, mean IoU_{3D} among matched true positives, median and 95th-percentile center error in millimetres, extent error for (w, h, d), precision, false positives per volume, and per-case detection count error. Classification endpoints are macro-F1 and class-specific sensitivity for R01–R04 among matched detections. Operating-point reporting includes the fold-specific selected threshold, calibration sensitivity,

and calibration FP/volume. Efficiency reporting distinguishes measured patch-level time and accelerator memory from complete reconstructed-volume latency, which was not established by the available profile.

All endpoints are reported overall and, where estimable, by vendor, region, single versus multiple missing sites, and geometry profile. Empty or structurally unsupported cells are labelled not estimable. In particular, the released data contain no Newtom R01 or R02 boxes; no Newtom-specific performance claim is made for those region classes.

4.22 Vendor robustness analyses

The primary subgroup analysis reports out-of-fold performance separately for Kavo, Meyer, and Newtom, alongside the maximum-minus-minimum vendor gap. A secondary strict leave-one-vendor-out (LOVO) domain-generalization experiment was completed after the original outer-test lock and is therefore explicitly post-lock exploratory. For each held-out vendor, both labeled and unlabeled scans from that vendor were excluded from training, and only the other two vendors supplied optimization, validation, calibration, and unlabeled data. The deterministic source-only partitions contained 49/10/10 optimization/validation/calibration cases when Kavo was held out, 17/3/3 when Meyer was held out, and 52/11/11 when Newtom was held out; the corresponding held-out targets contained 14, 60, and 9 cases. Five fixed repeats compared Pure supervised, vendor-adversarial, and full reliability-gated training, yielding 45 formal held-out model evaluations. Gated LOVO runs were initialized from the corresponding new source-only Pure checkpoint and used byte-identical Pure normalization statistics; vendor-adversarial runs fitted normalization from source optimization plus source unlabeled cases. No held-out-vendor case informed model selection.

For each trained LOVO model, checkpoint selection used source-validation pFROC only, with an exact tie retaining the earlier checkpoint. Operating thresholds were selected exclusively on the source calibration set from 1,001 thresholds on $[0, 1]$, maximizing sensitivity subject to FP/volume ≤ 1, then preferring lower FP/volume and finally the higher threshold. The threshold was frozen before held-out inference. Held-out pFROC was a separate threshold-swept exploratory metric using FP/volume budgets 0.5, 1, 2, and 4 and did not alter the operating threshold. Descriptive uncertainty used 2,000 case-level bootstrap replicates with seed 20260817; the same resampled cases were used across methods and the five stochastic repeats, and seed repeats were not treated as independent clinical units. No pairwise LOVO superiority test was prespecified. For Newtom, the small target sample and class-support differences preclude a strong universal-domain claim; the three-case source calibration partition in the Meyer-held-out setting is likewise a source of operating-point instability.

Vendor shortcut behavior is quantified by training a frozen-feature linear probe to predict vendor on development data and evaluating it on held-out cases. Reduced probe accuracy supports, but does not by itself prove, invariance; it must be interpreted jointly with maintained localization performance. Results are also compared before and after vendor balancing to distinguish adversarial alignment from sampling effects.

Table 17: Post-lock exploratory strict LOVO source/target design. Optimization, validation, and calibration columns contain labeled source-vendor cases only; source unlabeled scans exclude the held-out vendor.

Held-out vendor	Optimization	Validation	Calibration	Held-out cases	Source unlabeled	Held-out objects
Kavo	49	10	10	14	52	32
Meyer	17	3	3	60	41	68
Newtom	52	11	11	9	53	11

4.23 Repeated runs and checkpoint selection

Five fixed random seeds are used for every trainable method within every outer fold. Seeds control initialization, minibatch order, and augmentation, while split membership remains unchanged. For the executed Pure supervised, vendor-adversarial, and reliability-gated runs reported here, checkpoint selection maximized validation pFROC only; exact pFROC ties retained the earlier checkpoint. No center-error composite was used for checkpoint selection. Thresholds, stopping patience, and maximum epochs were otherwise fixed by the corresponding locked configuration.

The primary statistical unit remains the case, not the seed. Seed variability is summarized by mean, standard deviation, and range, but inferential intervals are obtained from paired case-level out-of-fold predictions. Hyperparameters are selected once from pilot development folds or by nested search and then frozen; the best configuration is not reselected separately for each outer-test fold. In post-lock LOVO, all model-selection rules were frozen before the first held-out-vendor access and no target-driven rerun was permitted.

4.24 Statistical analysis and uncertainty

For the original locked method contrasts, a paired hierarchical bootstrap resamples cases within vendor and repeats the complete metric calculation while retaining paired predictions from competing methods. The 2.5th and 97.5th percentiles form the 95% confidence interval. Center error is compared by paired bootstrap of the median; sensitivity differences use paired case-level resampling; and AP/FROC intervals are obtained by recomputing matching after each resample. As a robustness check, fold-stratified permutation tests exchange paired method labels within cases.

No superiority hypothesis is designated after inspection of the completed outer-test analysis. The original VR-SSL3D-versus-Supervised-3D hypothesis was not supported; newly initiated baseline, subgroup, calibration, ablation, and strict LOVO analyses are exploratory and are identified as such in tables and captions. For LOVO, no pairwise superiority test was prespecified; five-seed mean, sample SD, range, and case-level bootstrap 95% intervals are descriptive. For the family of four region-specific and three vendor-specific comparisons, false-discovery rate is controlled by Benjamini–Hochberg when formal p-values are reported. Effect sizes and confidence intervals remain primary; statistical significance alone is not interpreted as clinical relevance.

4.25 Missing labels, failed runs, and protocol deviations

Unreleased annotations are never imputed for evaluation. Outer-test metrics therefore concern only the released labeled cohort. A failed training run is rerun once with the identical seed after correcting an implementation or infrastructure fault; both the failure and reason are logged. Runs are not discarded because of poor performance. Any change to a locked split, endpoint, threshold, or baseline after outer-test access is reported as a protocol deviation and the affected analysis is labelled exploratory.

4.26 Locked reporting outputs

Before training, the following are checksum-locked: eligibility ledger, outer and inner manifests, duplicate registry, seeds, preprocessing configuration, hyperparameter search space, matching rules, endpoint definitions, and analysis script. Each completed run exports case-level predictions before aggregation. The Results section must include: (i) cohort flow and fold-balance table; (ii) primary baseline table with 95% intervals; (iii) vendor and region subgroup table; (iv) ablation table; (v) FROC curves; (vi) center-error and calibration plots; (vii) pseudo-box precision–coverage analysis; (viii) computational-cost table; and (ix) prespecified qualitative successes and failures. No performance number enters the manuscript without a traceable prediction file and run configuration.

4.27 Ethical approval and consent

This secondary analysis used a publicly released, de-identified CBCT dataset. Before submission, the authors must insert the approving ethics committee, protocol identifier, confirmation of compliance with relevant guidelines and the Declaration of Helsinki, and the documented informed-consent or waiver statement from the source dataset and responsible institution. No claim of exemption or waiver should be made without documentary verification.

Data availability

The source MORPH-CBCT dataset is described in the original data publication [25]. De-identified derived split manifests, aggregate audit tables, model configurations, analysis code, checksum manifests, and the case-level data necessary to reproduce the reported aggregate results will be deposited in a public repository before acceptance. Raw DICOM files remain subject to the source repository's access conditions and applicable privacy requirements. No directly identifying DICOM metadata are included in the manuscript or Supplementary Information.

References

[1] S. Agrawal et al.. The use of artificial intelligence in planning dental implant procedures: a systematic review. *Systematic Review* (2026). https://pubmed.ncbi.nlm.nih.gov/42187626/.

[2] M. Baumgartner, P. F. J"ager, F. Isensee, and K. H. Maier-Hein. nnDetection: a self-configuring method for medical object detection. *MICCAI* (2021). doi: https://doi.org/10.1007/978-3-030-87240-3_52.

[3] M. Bonfanti-Gris et al.. Deep learning for tooth detection and segmentation in panoramic radiographs: a systematic review and meta-analysis. *BMC Oral Health* (2025). https://pubmed.ncbi.nlm.nih.gov/40739210/.

[4] D. C. Castro, I. Walker, and B. Glocker. Causality matters in medical imaging. *Nature Communications* **11**, 3673 (2020). doi: https://doi.org/10.1038/s41467-020-17478-w.

[5] H. Chen et al.. A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films. *Scientific Reports* **9** (2019). doi: https://doi.org/10.1038/s41598-019-40414-y.

[6] X. Chen et al.. Semi-supervised semantic segmentation with cross pseudo supervision. *CVPR*, 2613–2622 (2021).

[7] G. S. Collins et al.. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024). doi: https://doi.org/10.1136/bmj-2023-078378.

[8] Z. Cui et al.. A two-stage framework for automated individual tooth segmentation and identification from cone-beam computed tomography images. *Medical Image Analysis* (2022). doi: https://doi.org/10.1016/j.media.2021.102220.

[9] Y. Ganin et al.. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* **17**(59), 1–35 (2016).

[10] R. Hamid et al.. Accuracy and time efficiency of artificial intelligence-driven tooth segmentation on CBCT images: a validation study using two implant planning software programs. *Clinical Oral Implants Research* (2025). https://pubmed.ncbi.nlm.nih.gov/40682303/.

[11] ImplantPlanNet Consortium. ImplantPlanNet: a deep learning framework for automatic implant planning from preoperative CBCT images. *Journal of Dentistry* (2026). https://pubmed.ncbi.nlm.nih.gov/42508797/.

[12] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**, 203–211 (2021). doi: https://doi.org/10.1038/s41592-020-01008-z.

[13] N. Jha et al.. Multiclass CBCT image segmentation for orthodontics with deep learning. *Journal of Dental Research* (2022). https://pubmed.ncbi.nlm.nih.gov/33783247/.

[14] M. Kennerley, J.-G. Wang, B. Veeravalli, and R. T. Tan. CAT: exploiting inter-class dynamics for domain adaptive object detection. *CVPR*, 16541–16550 (2024).

[15] E. Kim et al.. Classification of presence of missing teeth in each quadrant using deep learning artificial intelligence on panoramic radiographs of pediatric patients. *Journal of Clinical Pediatric Dentistry* (2024). doi: https://doi.org/10.22514/jocpd.2024.062.

[16] Y.-J. Li et al.. Cross-domain adaptive teacher for object detection. *CVPR*, 7581–7590 (2022).

[17] T.-Y. Lin et al.. Focal loss for dense object detection. *ICCV*, 2980–2988 (2017).

[18] X. Liu et al.. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine* **26**, 1364–1374 (2020). doi: https://doi.org/10.1038/s41591-020-1034-x.

[19] Y. Liu et al.. Novel AI-based tool for primary tooth segmentation on CBCT using convolutional neural networks: a validation study. *International Journal of Paediatric Dentistry* (2024). https://pubmed.ncbi.nlm.nih.gov/38769619/.

[20] Y.-C. Liu et al.. Unbiased teacher for semi-supervised object detection. *ICLR* (2021).

[21] X. Luo et al.. Efficient semi-supervised gross target volume segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis* **80**, 102517 (2022). doi: https://doi.org/10.1016/j.media.2022.102517.

[22] F. Mangano et al.. The role of artificial intelligence in implant dentistry: a systematic review. *Journal of Dentistry* (2025). https://pubmed.ncbi.nlm.nih.gov/40436717/.

[23] G. Miranda et al.. Emergence of artificial intelligence for automating cone-beam computed tomography-derived maxillary sinus imaging tasks: a systematic review. *Clinical Implant Dentistry and Related Research* (2024). https://pubmed.ncbi.nlm.nih.gov/38863306/.

[24] J. Mongan, L. Moy, and C. E. Kahn. Checklist for artificial intelligence in medical imaging: a guide for authors and reviewers. *Radiology: Artificial Intelligence* **2**(2), e200029 (2020). doi: https://doi.org/10.1148/ryai.2020200029.

[25] MORPH-CBCT Consortium. Adults' dental cone beam computed tomography images dataset for detecting and classifying missing teeth. *Scientific Data* (2026). https://pubmed.ncbi.nlm.nih.gov/41942491/. PMID: 41942491.

[26] OraSeg Consortium. An open deep learning-based framework and model for tooth instance segmentation in dental CBCT. *Clinical Oral Investigations* (2025). https://pubmed.ncbi.nlm.nih.gov/40996470/.

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *NeurIPS* (2015).

[28] M. Revilla-Leon et al.. Dental implant planning using artificial intelligence: a systematic review and meta-analysis. *Journal of Prosthetic Dentistry* (2024). https://pubmed.ncbi.nlm.nih.gov/38653687/.

[29] S. C. Rivera et al.. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine* **26**, 1351–1363 (2020). doi: https://doi.org/10.1038/s41591-020-1037-7.

[30] F. Schwendicke et al.. Trueness of artificial intelligence-driven CBCT tooth segmentation: a comparative validation ex vivo pilot study. *Journal of Dentistry* (2026). https://pubmed.ncbi.nlm.nih.gov/41796932/.

[31] K. Sohn et al.. FixMatch: simplifying semi-supervised learning with consistency and confidence. *NeurIPS* (2020).

[32] K. Sohn et al.. A simple semi-supervised learning framework for object detection. *arXiv:2005.04757* (2020). https://arxiv.org/abs/2005.04757.

[33] K. Stacke, G. Eilertsen, J. Unger, and C. Lundstr"om. A closer look at domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics* **25**(2), 325–336 (2021). doi: https://doi.org/10.1109/JBHI.2020.3001083.

[34] B. Sun and K. Saenko. Deep CORAL: correlation alignment for deep domain adaptation. *ECCV Workshops* (2016). doi: https://doi.org/10.1007/978-3-319-49409-8_35.

[35] Y. Tang, W. Chen, Y. Luo, and Y. Zhang. Humble teachers teach better students for semi-supervised object detection. *CVPR*, 3132–3141 (2021).

[36] TAPSeg Consortium. TAPSeg: an open-source deep learning tool for instance-level tooth and pulp segmentation in CBCT. *International Endodontic Journal* (2026). https://pubmed.ncbi.nlm.nih.gov/41865812/.

[37] A. Tarvainen and H. Valpola. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS* (2017).

[38] Z. Tian, C. Shen, H. Chen, and T. He. FCOS: fully convolutional one-stage object detection. *ICCV*, 9627–9636 (2019).

[39] ToothFairy2 Consortium. Multi-structure segmentation in CBCT volumes: the ToothFairy2 challenge. *Medical Image Analysis* (2026). https://pubmed.ncbi.nlm.nih.gov/42102640/.

[40] D. V. Tuzoff et al.. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofacial Radiology* **48**(4) (2019). doi: https://doi.org/10.1259/dmfr.20180051.

[41] S. Vinayahalingam et al.. Automated chart filing on panoramic radiographs using deep learning. *Journal of Dentistry* **115** (2021). doi: https://doi.org/10.1016/j.jdent.2021.103864.

[42] L. Wang et al.. A fully automated method for 3D individual tooth identification and segmentation in dental CBCT. *IEEE Transactions on Medical Imaging* (2021). https://pubmed.ncbi.nlm.nih.gov/34077356/.

[43] M. Xu et al.. End-to-end semi-supervised object detection with soft teacher. *ICCV*, 3060–3069 (2021).

[44] K. Yan, X. Wang, L. Lu, and R. M. Summers. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging* **5**(3), 036501 (2018). doi: https://doi.org/10.1117/1.JMI.5.3.036501.

[45] K. Yan et al.. Universal lesion detection by learning from multiple heterogeneous datasets. *MICCAI* (2019).

[46] Y. Yang, G. Sun, T. Zhang, R. Wang, and J. Su. Semi-supervised medical image segmentation via weak-to-strong perturbation consistency and edge-aware contrastive representation. *Medical Image Analysis* **101**, 103450 (2025). doi: https://doi.org/10.1016/j.media.2024.103450.

[47] L. Yu et al.. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. *MICCAI* (2019). doi: https://doi.org/10.1007/978-3-030-32245-8_66.

[48] Y. Yuan, X. Wang, X. Yang, and P.-A. Heng. Effective semi-supervised medical image segmentation with probabilistic representations and prototype learning. *IEEE Transactions on Medical Imaging* **44**(3), 1181–1193 (2025). doi: https://doi.org/10.1109/TMI.2024.3484166.

[49] J. R. Zech et al.. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLOS Medicine* **15**(11), e1002683 (2018). doi: https://doi.org/10.1371/journal.pmed.1002683.

[50] Q. Zeng et al.. Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging* **44**(1), 44–56 (2025). doi: https://doi.org/10.1109/TMI.2024.3429340.

[51] Z. Zhao et al.. Masked retraining teacher-student framework for domain adaptive object detection. *ICCV*, 19039–19049 (2023).

[52] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang. Domain generalization with MixStyle. *ICLR* (2021).

[53] X. Zhou, D. Wang, and P. Kr"ahenb"uhl. Objects as points. *arXiv:1904.07850* (2019).

Author contributions

H.A. contributed to conceptualization, methodology, software implementation, formal analysis, validation, visualization, and preparation of the original manuscript draft. F.Z. contributed to conceptualization, methodology, project administration, supervision, validation, and manuscript review and editing. A.T. contributed to clinical and radiological validation, interpretation of the findings, and manuscript review and editing. E.M. contributed to methodological review, supervision, validation, and manuscript review and editing. All authors reviewed and approved the final manuscript.

Funding

The authors received no specific funding for this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information accompanies this paper. Correspondence and requests for materials should be addressed to Faraneh Zarafshan (Fa.zarafshan@iau.ac.ir).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PAPER2ScientificReportsSupplementaryFINALv183.pdf