Ontology-Grounded Structured Prediction for Dental CBCT Reporting

Luca Lumetti[1], Mattia Di Bartolomeo[2], Arrigo Pellacani[3], Alexandre Anesi[1], Costantino Grana[1], and Federico Bolelli[1, ✉]

[1]University of Modena and Reggio Emilia, Italy
[2]Independent Researcher, Italy
[3]University Hospital of Modena, Italy

Abstract. We present a dataset and baseline for ontology-grounded structured prediction from dental Cone-Beam Computed Tomography (CBCT) volumes. Building on the public ToothFairy3 benchmark (532 volumes with expert-level segmentations), we contribute *(i)* a total of 893 free-text clinical reports for 529 publicly available CBCT volumes, *(ii)* their conversion into validated RDF/Turtle (Resource Description Framework) instances aligned with a clinician-designed OWL (Web Ontology Language) ontology spanning 13 finding types and multiple qualifier axes, and *(iii)* a strong baseline demonstrating the effectiveness of our setup and establishing a foundation for future work. We formulate CBCT reporting as a *three-stage* structured prediction problem—i.e., finding *detection*, anatomical *slot allocation*, and *property* prediction—and introduce a hierarchical evaluation suite of six clinically interpretable metrics that decouple detection, localization, and characterization. A baseline model using frozen multi-scale VoxTell features, a structure-indexed encoder, and ontology-driven prediction heads achieves strong results under 5-fold cross-validation, with stage-decoupled analysis identifying presence detection as the primary deployment bottleneck. Dataset, ontology, and code are publicly released.[1]

Keywords: Cone-Beam Computed Tomography · Report Generation · Medical Ontology · ToothFairy

1 Introduction

Cone-Beam Computed Tomography (CBCT) is central to dental and maxillofacial practice, supporting implant planning, surgery, and pathology assessment by providing volumetric visualization of teeth, jawbones, canals, and adjacent structures. In routine workflows, clinicians translate CBCT observations into reports that specify *what* is present, *where* it is located (often tooth-specific), and *how* it should be characterized (e.g., severity, morphology, treatment quality). Automating reporting could reduce clinical workload and enable screening, but

[1] https://github.com/AImageLab-zip/CBCT-Report

✉ Corresponding author: federico.bolelli@unimore.it

requires models that jointly handle multiple anatomical structures and compositional clinical findings.

Most CBCT benchmarks emphasize segmentation, focusing on structures such as teeth, jaws, inferior alveolar canals, and pulp [8,11,12,16,18,20,23]. The ToothFairy challenge series [5,6,7,17] has advanced multi-structure maxillofacial segmentation, culminating in ToothFairy3 with 532 public CBCT scans and 50 private test scans for long-term benchmarking [19], each accompanied by 77 different accurate segmentation classes. However, moving from segmentation to finding detection and reporting introduces challenges: findings are heterogeneous and imbalanced, must be grounded to an anatomical region or structure (e.g., ISO 3950/FDI tooth identifiers [9]), and require clinically relevant qualifiers.

A key bottleneck for generative reporting is evaluation. Standard free-text captioning metrics (e.g., BLEU [21], ROUGE [13], METEOR [2]) measure surface similarity and can correlate poorly with clinical correctness, missing hallucinations, omissions, or wrong localization [1,3,4,14,26]. Clinically informed alternatives such as entity/relation extraction (e.g., RadGraph [10]) or LLM-based factuality scoring (e.g., RadFact [3]) remain domain- and schema-dependent; for dental CBCT, tooth-indexed multi-instance findings and property axes are rarely standardized.

We therefore formulate CBCT reporting as ontology-grounded[2] structured prediction rather than free-text generation. We introduce a resource derived from the publicly available ToothFairy3 cases, consisting of *(i)* free-form clinical reports authored by clinical staff, and *(ii)* their conversion into validated ontology instances. The central representation is a clinician-authored OWL/RDF (Web Ontology Language/Resource Description Framework) ontology defining 13 finding types and orthogonal qualifier axes (e.g., severity, morphology, orientation), serialized in Turtle (Terse RDF Triple Language) [25,24]. A companion SHACL (Shapes Constraint Language) specification constrains applicable properties, cardinalities, and value ranges, enabling automated validation and schema-driven model head configuration.

Contributions. In summary, the paper provides *(i)* 893 free-text reports for 529 ToothFairy3 cases and validated ontology instances, enabling learning and evaluation beyond segmentation; *(ii)* an OWL/RDF schema spanning 13 finding types and multiple qualifier axes, with ad-hoc SHACL shapes for validation and model head auto-configuration; and *(iii)* a multi-stage framework predicting finding *presence*, *anatomical slots*, and *properties*, with code for reproducibility.

2 The Proposed Dataset and Ground-Truth Construction

We build on the ToothFairy3 benchmark, which provides 532 public CBCT volumes of the maxillofacial region, each paired with expert ground-truth multi-

[2] An ontology is a clinician-defined, machine-readable model of the reporting domain. It lists relevant concepts (e.g., findings and anatomical entities) and restricts how they can be combined using relations and attributes (e.g., severity and morphology).

Algorithm 1: Instance excerpt for `Patient_3F_004__Finding_1`

rdfs:comment

```
Mandibular canal with a predominantly lingual course, in close
 relationship with the roots of dental elements 37 and 38.
```

rdf:properties

```
:findingType → :MandibularCanalCourse
:onSide → :Left
:hasCanalTrajectory → :PrevalentlyLingual
:hasCanalToothContact → :IntimateContact
:hasToothId → "37", "38"
```

structure segmentation masks (77 segmentation classes) and tooth-level localization consistent with ISO 3950/FDI identifiers.
Our reports were authored by 2 radiologists and 6 maxillofacial surgeons with 5–20 years of experience in CBCT interpretation, who examined the anonymized CBCT volumes. No standardized reporting template was imposed; reports were primarily free-text narratives that followed routine clinical practice, typically including systematic evaluation of the dentition, periapical regions, and bone structures. In total, the generated dataset comprises *893 free-text reports* across 529 of the 532 cases (one or two per patient); three patients were excluded due to image-quality issues, making report authoring unreliable.

We convert free-text CBCT reports into structured supervision targets (ontology instances) that are directly learnable and automatically checkable. Concretely, each report becomes a patient-level RDF graph composed of atomic *finding* instances that encode *what* was observed (a *finding type*) and *where* it applies (i.e., tooth, side, or quadrant). Each finding can also include standardized *qualifier properties* that describe *how* it should be characterized (e.g., severity and morphology). To ensure consistency, we validate all graphs against SHACL constraints and perform clinician-guided iterative corrections. An example of a finding extracted from an ontology instance is provided in Alg. 1. We suggest the reader explore a complete ontology instance for a broader understanding.[3]

Target Representation: Ontology-grounded Findings (*what* and *where*). Findings are the smallest reportable CBCT observations in our target space. In our scenario, each finding has *exactly one* finding type from a *closed set of 13* clinically validated categories. Anatomical grounding is encoded via the corresponding slot (*where*). Bilateral observations are always split into two independent finding instances, i.e., left and right. Tooth identifiers support multi-tooth entities, e.g., a bridge spanning teeth 14–16 carries three `:hasToothId` assertions, each activating its own slot bit independently.

Anatomical Properties (*how*). They enforce single-valued semantics (e.g., `:hasSeverity` ∈ {`Mild`, `Moderate`, `Severe`}; `:hasRadiodensity` ∈ {`Radiolu-`

[3] https://github.com/AImageLab-zip/CBCT-Report/blob/main/data/ttl_reports/ToothFairy3F_001.ttl

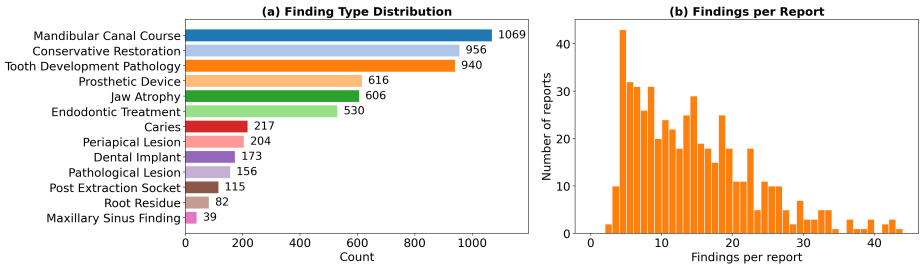

Fig. 1. Distribution of the number of extracted findings per patient and the distribution of each finding type in the dataset.

`cent`, `Radiopaque`, `MixedDensity`}), or multi-label assertions (e.g., `:hasMorphology`, `:hasPeriodontalFeature`).

2.1 SHACL Validation and Automated Extraction

SHACL (Shapes Constraint Language) is a standard used to validate RDF graph data against a set of conditions or constraints. Our SHACL core shape requires exactly one `:findingType`, constrains qualifier value ranges, and validates tooth IDs using an FDI regex. Additional shapes restrict qualifiers by finding type and specific conditional requirements.

In our context, ground-truth instances are produced from free-text reports using GPT-5.2. For each patient, the prompt includes: *(i)* the OWL/RDF ontology schema defined by clinicians, *(ii)* the raw report text (both reports when available) with instructions to consolidate duplicates, *(iii)* three few-shot examples authored by clinicians, and *(iv)* explicit extraction rules. The model is tasked to output pure Turtle code, i.e., a syntax for defining RDF graphs. Outputs are parsed with `rdflib` and validated with `pySHACL`. Each extracted finding includes an `rdfs:comment` that quotes the source phrase, enabling targeted auditing. The extracted ontology instances (our labels) are released alongside free-text reports and the final validated ontology instances, enabling future researchers to use the labels without re-running the LLM extraction pipeline.

Supervised Iterative Refinement. We employ an iterative and interactive extraction process: each SHACL violation (i.e., an output that is non-compliant with the defined rules/shapes) was reviewed by a dental clinician who either corrected a report typo, updated the ontology constraint, or manually authored the instance for that patient. When systematic errors were identified, the prompt and few-shot examples were updated, and affected cases were re-extracted. Iterations continued until no extraction errors remained.

After the SHACL-based structural validation process, all generated ontology instances were further manually verified by clinicians, leveraging the free-text report excerpt from which the finding was extracted. Notably, the final structured graphs represent *clinician-validated annotations rather than raw LLM outputs.*

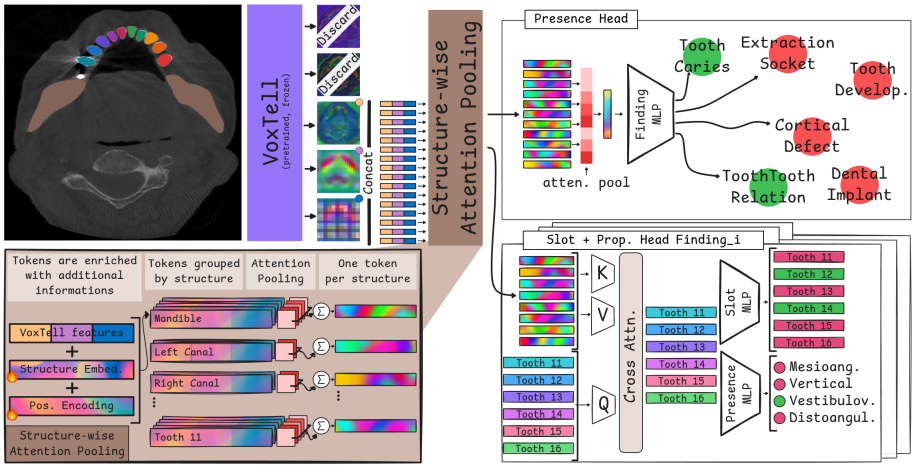

Fig. 2. Overview of the proposed pipeline (Sec. 3). Frozen VoxTell features are tokenized per anatomical structure and pooled into one embedding per structure. A presence head detects which finding types are active; per-finding-type slot heads localize each finding to its anatomical index (tooth/side/quadrant) and predict its qualifiers.

Dataset Statistics and Label Distribution. Across successfully parsed cases, the extraction pipeline yields a total of *5,703* finding instances over the *13* finding types, exhibiting strong class imbalance consistent with clinical prevalence. Fig. 1 shows all selected types and their distribution.

Clinical Agreement. To quantify human agreement, we separately applied the above-described free-text-to-ontology automatic conversion pipeline to the subset of 358 patients for whom at least two independent reports were available. We refer the reader to Sec. 4, where the evaluation metrics and the corresponding agreement scores are reported.

3 Methods

The proposed ontology allows us to reframe the free-text report generation task into a "simpler" classification problem. To this aim, we develop and propose a model that maps a CBCT volume and its multi-structure segmentation to an ontology-grounded structured report through four pipeline stages: *(i)* frozen multi-scale feature extraction, *(ii)* segmentation-guided regional tokenization, *(iii)* a trainable encoder producing one embedding per anatomical structure, and *(iv)* ontology-driven prediction heads for finding presence (*what*), slot allocation (*where*), and property prediction (*how*). The pipeline is depicted in Fig. 2.

Feature Extraction and Tokenization. We adopt VoxTell [22] as a pretrained and frozen 3D encoder. Features from encoder levels 3, 4, and 5 are upsampled to the level-3 spatial grid and concatenated along the channel dimension (C=896),

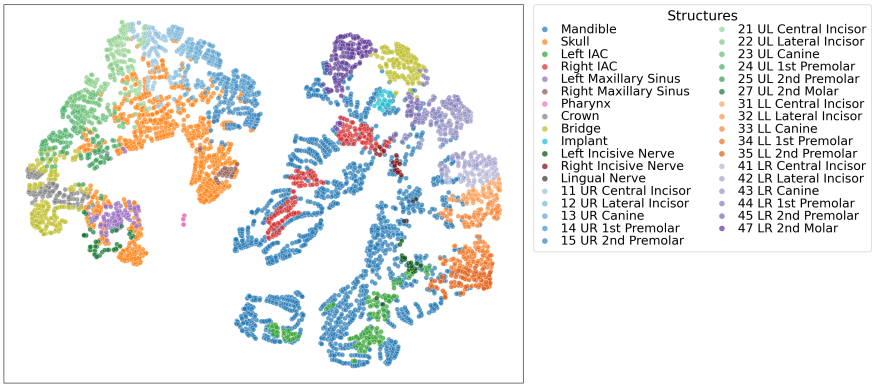

Fig. 3. t-SNE applied to extracted tokens after applying max-pooling to the largest structures. Features used are the concatenation of VoxTell Stages 3, 4, and 5.

combining coarse semantic content (level 5) with finer spatial detail (level 3). The ground-truth segmentation mask from the ToothFairy3 dataset is downsampled to this grid via adaptive max-pooling to preserve structure labels.

This provides typically ∼10,000 structure-aligned tokens per patient. Fig. 3 shows a t-SNE visualization of extracted structure-level tokens, confirming that embeddings form semantically coherent clusters by anatomical category.

Structure Encoder. Tokens are then projected to d=256, enriched with a learned structure-type embedding (8 coarse anatomical categories) and a spatial positional encoding, then grouped by structure ID. Within each group, additive attention pooling collapses the tokens into a single structure embedding:

$$\mathbf{s}_\ell = \sum_{n \in \mathcal{I}_\ell} \alpha_{\ell,n}\, \tilde{\mathbf{x}}_n, \quad \alpha_{\ell,n} = \mathrm{softmax}_{n \in \mathcal{I}_\ell}\big(\mathbf{v}^\top \tanh(\mathbf{W}\tilde{\mathbf{x}}_n)\big). \tag{1}$$

Here, ℓ indexes an anatomical structure and $\mathcal{I}_\ell$ denotes the set of token indices belonging to structure ℓ; $\tilde{\mathbf{x}}_n \in \mathbb{R}^d$ is the embedded feature vector of token n. The pooled structure representation $\mathbf{s}_\ell \in \mathbb{R}^d$ is obtained as an attention-weighted sum of its tokens, where $\alpha_{\ell,n}$ is the normalized attention weight computed via additive attention: $\mathbf{W}$ is a learnable projection matrix, $\tanh(\cdot)$ is applied element-wise, and $\mathbf{v}$ is a learnable vector producing a scalar score per token. The softmax is taken over $n \in \mathcal{I}_\ell$ so that $\sum_{n \in \mathcal{I}_\ell} \alpha_{\ell,n} = 1$. The resulting set of tokens is defined as $\mathbf{H} \in \mathbb{R}^{B \times S \times 256}$.

Ontology-Driven Prediction Heads. Head configuration is derived automatically from our SHACL schema at each stage.

Stage 1 — finding presence (what). A global attention pool over $\mathbf{H}$ produces a patient-level embedding, from which a small MLP predicts 13 sigmoid logits.

Stage 2 — slot allocation (where). For each finding type f with K_f ontology-defined slots, slot embeddings are computed via cross-attention to $\mathbf{H}$ and a

binary activation is predicted per slot:

$$\mathbf{Z}_f = \text{CrossAttn}(\mathbf{Q}_f, \mathbf{H}, \mathbf{H}) \in \mathbb{R}^{B\times K_f\times 256}, \qquad \hat{y}^{\text{slot}}_{f,k} = \sigma(\mathbf{w}_f^\top \mathbf{Z}_{f,k} + b_f). \quad (2)$$

The slot queries $\mathbf{Q}_f$ follow the ontology indexing scheme: tooth-indexed finding types use K_f=32 slots (one per FDI tooth 11–48) with queries initialized from the corresponding tooth embeddings in $\mathbf{H}$. Side-indexed types (K_f=2) and quadrant-indexed types (K_f=4) use fully learned queries.

Stage 3 — property prediction (how). For each active (f, k) slot, a per-property MLP maps $\mathbf{Z}_{f,k}$ to qualifier predictions: softmax for single-choice categorical properties, sigmoid for multi-label. During training, losses are applied only to ground-truth active slots.

Training Objective. We minimize a weighted sum of per-stage losses:

$$\mathcal{L} = \mathcal{L}_{\text{pres}} + \mathcal{L}_{\text{slot}} + \mathcal{L}_{\text{prop}}. \quad (3)$$

$\mathcal{L}_{\text{pres}}$ and $\mathcal{L}_{\text{slot}}$ use focal-BCE (γ=0.5) with per-class positive weights; $\mathcal{L}_{\text{slot}}$ additionally applies label smoothing (ϵ=0.1). $\mathcal{L}_{\text{prop}}$ uses cross-entropy for categorical and BCE for multi-label properties, both with label smoothing (ϵ=0.1).

4 Experiments

Evaluation Protocol. A clinically useful structured report must simultaneously satisfy *detection*, *localization*, and *characterization* axes. To this aim, we evaluate performance at four levels, aligned with the staged task:

Finding-Type Detection (pres-F1). Macro F1 over the 13 finding types, ignoring localization and properties.

Instance Localization (inst-F1). Macro F1 over ⟨finding type, slot⟩ pairs. A predicted slot activation is a true positive only if both the finding type and the anatomical slot match a ground-truth target exactly.

Characterization Accuracy (char-acc). Mean per-property accuracy over true-positive localized instances, isolating characterization from upstream errors.

Patient-Level Report Quality (rep-F1). Per-patient F1 between the predicted and ground-truth sets of ⟨finding type, slot⟩ pairs, averaged across patients.

To decouple pipeline stages for ablation, we also report *slot-F1* (slot-activation F1 conditioned on ground-truth present finding types) and *prop-F1* (property F1 conditioned on ground-truth active slots). These GT-conditioned metrics isolate localization and characterization quality independently of upstream errors.

Training Setup. We use all 529 ToothFairy3 volumes with successfully extracted ground-truth. A fixed held-out test set (20%, ∼107 volumes) is isolated via stratified sampling over finding types, slots, and properties; the remaining 80% (∼425 volumes) is partitioned into 5 folds, stratified by finding-type prevalence. Each fold trains on four-fifths of the pool (∼340 samples). We report means over folds. Training uses AdamW [15] with cosine LR schedule (η_{peak}=10^{-3},

Table 1. Effect of segmentation source on full-model performance (5-fold CV means ± std, fixed held-out test set). GT segmentations use ToothFairy3 masks; predicted segmentations are obtained from an nnU-Net trained on the same split.

Segmentation	pres-F1	slot-F1	prop-F1	inst-F1	char-acc	rep-F1
nnU-Net	71.2 ± 0.6	51.7 ± 2.2	50.1 ± 0.6	43.3 ± 0.9	51.0 ± 0.6	49.8 ± 1.7
GT	72.0 ± 1.0	53.4 ± 1.1	61.2 ± 1.5	46.7 ± 0.8	85.4 ± 0.7	53.8 ± 0.9

Table 2. Effect of VoxTell encoder level on full-model performance (5-fold CV means). Single-level inputs use features of that level only; concat uses levels 3+4+5.

Features	pres-F1	slot-F1	prop-F1	inst-F1	char-acc	rep-F1
Lv. 3	69.4 ± 1.8	52.4 ± 1.4	58.6 ± 1.7	44.1 ± 1.3	84.2 ± 0.9	51.1 ± 2.2
Lv. 4	70.1 ± 1.5	52.8 ± 1.2	59.4 ± 1.6	44.9 ± 1.2	84.6 ± 0.8	51.9 ± 1.9
Lv. 5	60.1 ± 3.1	48.7 ± 2.8	53.6 ± 2.2	37.2 ± 1.9	80.7 ± 1.5	43.1 ± 4.1
Concat	$\mathbf{72.0 \pm 1.3}$	$\mathbf{53.4 \pm 1.1}$	$\mathbf{61.2 \pm 1.5}$	$\mathbf{46.7 \pm 0.8}$	$\mathbf{85.4 \pm 0.7}$	$\mathbf{53.8 \pm 0.9}$

$\eta_{\min}$=10^{-5}), batch size 32, dropout p=0.2, and early stopping on validation pres-F1 (patience 50, max 150 epochs). Post-hoc per-class threshold calibration is applied on the validation fold before reporting test metrics.

Segmentation Sensitivity. Tab. 1 reports model performance when leveraging automatic segmentation produced by an nnU-Net trained on the same split *vs* the GT segmentation masks. The small gap between these two variations (pres-F1: 71.2 *vs.* 72.0; rep-F1: 49.8 *vs.* 53.8) indicates that the pipeline is robust to imperfect structure delineation and quantifies the contribution of segmentation quality to reporting performance.

Stage-Decoupling Analysis. In the same table (Tab. 1), distinct behaviors at each prediction stage are reported. Under GT-conditioned evaluation, slot-F1 = 53.4 and prop-F1 = 61.2 confirm that, given correct upstream inputs, both localization and qualifier prediction perform reliably. In a fully-predicted evaluation (segmentation predicted), inst-F1 drops to 43.3 due to cascading detection errors. Notably, with GT segmentation, char-acc is high (85.4): if a finding is correctly detected and localized, its clinical properties are predicted with high fidelity. This identifies *presence detection as the primary deployment bottleneck*, not characterization.

Multiscale Feature Ablation. Tab. 2 compares individual VoxTell encoder levels against the full multi-scale concatenation. Level 5 alone performs worst: at 8×16×16 resolution, structure-masked pooling collapses the localization signal. Levels 3 and 4 retain finer spatial detail and perform substantially better. Concatenating all three levels yields the best results, confirming that high-level semantic content and fine spatial resolution are complementary.

Clinical Agreement. Human agreement is high overall for finding detection (pres-F1 0.81±0.16) and characterization (char-acc. 0.83±0.13) and remains sub-

stantial for structured report matching (rep-F1 0.67±0.21) and instance localization (inst-F1 0.70±0.20).

5 Conclusion

We introduce the first ontology-grounded CBCT report resource, with 893 clinical reports and clinician-validated structured labels spanning 13 finding types. We also propose a hierarchical metric suite that decouples detection, localization, and characterization, replacing surface-level text similarity with clinically interpretable scores, and a lightweight ontology-driven baseline. Overall, the released dataset, ontology, metrics, and baseline provide a rigorous foundation for moving dental CBCT AI beyond segmentation toward clinically actionable interpretation.

Acknowledgments. This project has received funding from Fondazione di Modena, through FAR 2024 (E93C24002080007).

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

References

1. Babar, Z., van Laarhoven, T., Zanzotto, F.M., Marchiori, E.: Evaluating diagnostic content of AI-generated radiology reports of chest X-rays. Artificial Intelligence in Medicine **116** (2021)
2. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (2005)
3. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., Meissen, F., Ranjit, M., Srivastav, S., Gong, J., Codella, N.C.F., Falck, F., Oktay, O., Lungren, M.P., Wetscherek, M.T., Alvarez-Valle, J., Hyland, S.L.: MAIRA-2: Grounded Radiology Report Generation. arXiv preprint arXiv:2406.04449 (2024)
4. Boag, W., Kané, H., Rawat, S., Wei, J., Goehler, A.: A Pilot Study in Surveying Clinical Judgments to Evaluate Radiology Report Generation. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021)
5. Bolelli, F., Lumetti, L., van Nistelrooij, N., Vinayahalingam, S., Di Bartolomeo, M., Marchesini, K., Pellacani, A., Candeloro, E., Rosati, G., Xi, T., Isensee, F., Kirchhoff, Y., Krämer, L., Rokuss, M., Ulrich, C., Maier-Hein, K., Jiang, Y., Liu, Y., Wang, L., Wang, H., Chen, S., Cui, Z., Shi, P., Pan, Z., Liang, X., Ma, Q., Konukoglu, E., Wodzinski, M., Müller, H., Mai, H., Dang, X., Bhandary, S., Grosu, R., Bergé, S., Anesi, A., Grana, C.: Multi-Structure Segmentation in CBCT Volumes: the ToothFairy2 Challenge. Medical Image Analysis (2026)
6. Bolelli, F., Lumetti, L., Vinayahalingam, S., Di Bartolomeo, M., Pellacani, A., Marchesini, K., van Nistelrooij, N., van Lierop, P., Xi, T., Liu, Y., Xin, R., Yang, T., Wang, L., Wang, H., Xu, C., Cui, Z., Wodzinski, M., Müller, H., Kirchhoff, Y., Rokuss, M., Maier-Hein, K., Han, J., Kim, W., Ahn, H.G., Szczepański, T., Grzeszczyk, M.K., Korzeniowski, P., Caselles Ballester, V., Paolo Burgos-Artizzu, X., Prados Carrasco, F., Bergé, S., van Ginneken, B., Anesi, A., Grana, C.: Segmenting the Inferior Alveolar Canal in CBCT Volumes: the ToothFairy Challenge. IEEE Transactions on Medical Imaging (2024)
7. Bolelli, F., Marchesini, K., van Nistelrooij, N., Lumetti, L., Pipoli, V., Ficarra, E., Vinayahalingam, S., Grana, C.: Segmenting Maxillofacial Structures in CBCT Volumes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
8. Gamal, M., Baraka, M., Torki, M.: Automatic Mandibular Semantic Segmentation of Teeth Pulp Cavity and Root Canals, and Inferior Alveolar Nerve on Pulpy3D Dataset. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 (2024)
9. ISO: Dentistry - Designation system for teeth and areas of the oral cavity. https://www.iso.org/standard/68292.html (2016), accessed: 2024-11-11
10. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., Langlotz, C.P., Rajpurkar, P.: RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. arXiv preprint arXiv:2106.14463 (2021)
11. Jaskari, J., Sahlsten, J., Järnstedt, J., Mehtonen, H., Karhu, K., Sundqvist, O., Hietanen, A., Varjonen, V., Mattila, V., Kaski, K.: Deep Learning Method for Mandibular Canal Segmentation in Dental Cone Beam Computed Tomography Volumes. Scientific Reports **10**(1) (2020)

12. Lahoud, P., EzEldeen, M., Beznik, T., Willems, H., Leite, A., Van Gerven, A., Jacobs, R.: Artificial Intelligence for Fast and Accurate 3-Dimensional Tooth Segmentation on Cone-beam Computed Tomography. Journal of Endodontics **47**(5) (2021)
13. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out (2004)
14. Liu, Y., Wang, Z., Li, Y., Liang, X., Liu, L., Wang, L., Zhou, L.: MRScore: Evaluating Medical Report with LLM-based Reward System. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 (2024)
15. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. arXiv preprint arXiv:1711.05101 (2017)
16. Lumetti, L., Marchesini, K., Pipoli, V., Ficarra, E., Grana, C., Bolelli, F.: Taming Mambas for 3D Medical Image Segmentation. IEEE Access (2025)
17. Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., Grana, C.: Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. IEEE Access **12** (2024)
18. Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., Grana, C.: Location Matters: Harnessing Spatial Information to Enhance the Segmentation of the Inferior Alveolar Canal in CBCTs. In: 27th International Conference on Pattern Recognition (ICPR) (2024)
19. Lumetti, L., Tan, Z.Q., Borghi, L., van Nistelrooij, N., Rosati, G., Addison, O., Li, Y., Vinayahalingam, S., Grana, C., Bolelli, F.: ToothFairy3: Scaling CBCT Maxillofacial Segmentation to 77 Classes with U-Mamba2. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2026 (2026)
20. van Nistelrooij, N., Krämer, L., Kempers, S., Beyer, M., Bolelli, F., Xi, T., Bergé, S., Heiland, M., Maier-Hein, K.H., Vinayahalingam, S., Isensee, F.: ToothSeg: Robust Tooth Instance Segmentation and Numbering in CBCT using Deep Learning and Self-Correction. IEEE Journal of Biomedical and Health Informatics (2025)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (2002)
22. Rokuss, M., Langenberg, M., Kirchhoff, Y., Isensee, F., Hamm, B., Ulrich, C., Regnery, S., Bauer, L., Katsigiannopulos, E., Norajitra, T., et al.: VoxTell: Free-Text Promptable Universal 3D Medical Image Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)
23. Usman, M., Rehman, A., Saleem, A.M., Jawaid, R., Byon, S.S., Kim, S.H., Lee, B.D., Heo, M.S., Shin, Y.G.: Dual-Stage Deeply Supervised Attention-Based Convolutional Neural Networks for Mandibular Canal Segmentation in CBCT Scans. Sensors **22**(24) (2022)
24. W3C OWL Working Group: OWL 2 Web Ontology Language Document Overview (2009)
25. World Wide Web Consortium: RDF 1.1 Turtle: terse RDF triple language (2014)
26. Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., et al.: Evaluating progress in automatic chest X-ray radiology report generation. Patterns **4**(9) (2023)