# Automatic Tooth Numbering in Panoramic X-Rays Using YOLOv8 and MaskR-CNN

Murilo Salem
Federal University of Pelotas (UFPel)
Pelotas, Brazil
mcsalem@inf.ufpel.edu.br

Rafael Grasel
Federal University of Rio Grande do Sul (UFRGS)
Porto Alegre, Brazil
rgrasel0@gmail.com

Daniel Barretos
Federal University of Pelotas (UFPel)
Pelotas, Brazil
dhspbarretos@inf.ufpel.edu.br

Anderson Ferrugem
Federal University of Pelotas (UFPel)
Pelotas, Brazil
ferrugem@inf.ufpel.edu.br

Alessandro Bof
Federal University of Pampa (UNIPAMPA)
Alegrete, Brazil
alessandrooliveira@unipampa.edu.br

Maurício Braga de Paula
Federal University of Pelotas (UFPel)
Pelotas, Brazil
maubrapa@gmail.com

Dante Augusto Couto Barone
Federal University of Rio Grande do Sul (UFRGS)
Porto Alegre, Brazil
dante.barone@gmail.com

Jonas Almeida Rodrigues
Federal University of Rio Grande do Sul (UFRGS)
Porto Alegre, Brazil
jonasrodrigues@ufrgs.br

## ABSTRACT

This paper presents an extended comparative evaluation of Mask R-CNN and YOLOv8 for automatic tooth counting and FDI numbering in panoramic radiographs. The study uses a private clinical dataset with 410 anonymized images and 53 dental classes, covering permanent teeth, deciduous teeth, and a supranumerary class. The manuscript preserves the original inter-architecture comparison between a two-stage Mask R-CNN reference and one-stage YOLO-based segmentation, while extending the reproducible experimental evidence on the YOLOv8 branch through model-scale comparison, expanded augmentation, calibrated post-processing, class-group analysis, five-fold robustness assessment, and a count-oriented audit. On the fixed validation split (328 training images, 82 validation images), YOLOv8s-seg achieved the best mAP50 for mask segmentation (0.9003), while YOLOv8m-seg achieved the best mAP50–95 (0.6673). A five-fold validation with the medium protocol produced mean scores of $0.8770 \pm 0.0132$ for mAP50 and $0.6374 \pm 0.0165$ for mAP50–95. The count-oriented audit of the post-processed medium configuration yielded a mean absolute counting error of 2.49 teeth per image, exact count agreement in 23.2% of validation cases, and dentition-type agreement in 81 of 82 images. The results preserve the original comparative message while strengthening the evidence that the small and medium YOLOv8 configurations provide the strongest practical trade-off, whereas mixed dentition, rare classes, and exact tooth counts remain the main open difficulties.

## KEYWORDS

Dental radiography; Tooth counting; Tooth numbering; Panoramic X-ray; Comparative evaluation; Mask R-CNN; YOLOv8; Instance segmentation; FDI; Medical imaging.

## 1 INTRODUCTION

Artificial intelligence has become increasingly relevant in dental imaging, especially for tasks that depend on rapid and consistent interpretation of radiographs [7, 9]. Panoramic radiographs are particularly attractive for this purpose because they provide a single-view summary of the maxillary and mandibular arches, supporting screening, treatment planning, and documentation workflows [5, 13].

Automatic tooth counting and numbering are central subtasks in this context. A model that correctly identifies individual teeth and assigns FDI labels can support structured reporting, pre-annotation, longitudinal follow-up, and downstream diagnostic systems. The problem remains difficult in practice because panoramic exams contain geometric distortion, overlapping structures, mixed dentition, restorations, partially erupted teeth, and class imbalance across the dental arch.

Recent dental AI benchmarks have shown that detection and segmentation quality depend not only on the backbone architecture, but also on label granularity, training protocol, and clinical diversity of the dataset [2, 4]. In this work, we preserve the original comparative question: how do a two-stage Mask R-CNN reference and YOLO-based segmentation configurations behave for FDI-aware tooth counting and numbering on a real private panoramic dataset?

This extended version preserves the same dataset, annotation strategy, clinical task, and comparison core, while strengthening the experimental evidence with additional quantitative and qualitative analyses. Rather than reframing the study as a new model proposal, we treat the added validation as a deeper assessment of the compared configurations. The new end-to-end reruns reported here concentrate on the YOLOv8 branch, which was the strongest and most reproducible line in the current environment, while Mask

R-CNN remains explicit in the comparative framing and representative qualitative example inherited from the original study.

The main contributions of this paper are:

- preservation of the original comparative framing between Mask R-CNN and YOLO-based segmentation for FDI-aware tooth numbering on panoramic radiographs;
- a comparative evaluation of YOLOv8n-seg, YOLOv8s-seg, and YOLOv8m-seg on the 53-class panoramic dataset, expanding the strongest branch of the original study;
- a supplementary assessment of expanded augmentation and calibrated post-processing applied after the core model comparison;
- an extended validation protocol with class-group analysis, five-fold robustness, and a count-oriented audit;
- a discussion of practical limitations and validity threats for clinical deployment in a single-GPU research setting.

## 2 CONVOLUTIONAL NEURAL NETWORKS FOR DENTAL RADIOGRAPHIC ANALYSIS

Convolutional neural networks are widely used in dental imaging because they can learn hierarchical representations directly from radiographic data. In panoramic radiographs, CNN-based systems have been applied to detection, numbering, caries analysis, and broader diagnostic support tasks [1, 12]. These studies motivate comparative evaluation rather than commitment to a single recipe, since robustness, efficiency, and mask quality can vary substantially across configurations.

### 2.1 Compared Segmentation Configurations

YOLO treats object detection as a single-stage prediction problem, making it attractive when throughput and deployment simplicity matter [6]. YOLOv8 extends this family with anchor-free heads and an instance-segmentation branch that predicts class labels, masks, and localization jointly [11]. For tooth numbering, this is useful because each predicted instance can be mapped directly to an FDI class, and the number of detected instances per image naturally defines a tooth count.

In the present study we evaluated three YOLOv8 segmentation scales: nano, small, and medium. The comparison is relevant because the dental setting imposes a trade-off between capacity and robustness. Smaller models are attractive for fast pre-annotation, while larger models may better resolve overlapping or partially erupted teeth.

Mask R-CNN remains the original two-stage reference architecture for the manuscript [3]. Its region-first design is historically relevant for instance segmentation because it separates proposal generation from mask prediction, which can be appealing when individual object delineation matters. In this revised version, Mask R-CNN anchors the original comparative perspective, while the new reproducible experiments expand the YOLOv8 branch that was fully available and operational in the current codebase.

**Table 1: Dataset statistics for the fixed train/validation split.**

| Split | Images | Instances | Mean teeth/img | Median |
|---|---|---|---|---|
| Train | 328 | 12,064 | 36.78 | 37 |
| Validation | 82 | 3,062 | 37.34 | 38 |

### 2.2 Supplementary Extensions for the Medium Configuration

Beyond model scale, we evaluated two practical extensions of the baseline YOLOv8m-seg pipeline. The first extension adds a broader training recipe with intensity-based augmentation, light geometric perturbation, and weighted sampling to partially compensate for rare dental classes. The second extension calibrates the operating point at inference time using confidence threshold, IoU threshold, and a minimum area ratio filter to suppress anatomically implausible masks.

These additions were tested as ablations rather than assumptions of improvement. This distinction is important because curated medical datasets do not always benefit from stronger augmentation if the added variability departs from clinically plausible structure.

## 3 CHALLENGES IN APPLYING AI TO DENTAL RADIOGRAPHY

Panoramic radiographs remain challenging for automatic analysis. Superposition of structures, variable eruption stage, metallic artifacts, truncated fields, and asymmetries along the arches can degrade both localization and segmentation quality [12, 13]. Mixed dentition is especially difficult because deciduous and permanent teeth coexist, producing crowded regions with small objects and ambiguous spacing.

Rare classes create an additional challenge. In our dataset, the supranumerary class appeared only 13 times in the training split, which is far below the support available for the most frequent permanent classes. In such a regime, overall mAP can remain high while clinically relevant edge cases still fail consistently.

## 4 MATERIALS AND METHODS

### 4.1 Dataset Description and Characterization

The study uses a clinical dataset of 410 anonymized panoramic radiographs, curated under ethical and privacy constraints consistent with current recommendations for dental AI research [8]. The task is framed as 53-class instance segmentation with FDI-aware labels: 32 permanent teeth, 20 deciduous teeth, and one supranumerary class. A public Kaggle release of the dataset is available online [10].

The official split used throughout the project contains 328 training images and 82 validation images. The training split contains 12,064 annotated tooth instances, while the validation split contains 3,062 instances. All 53 classes are represented in validation. The mean number of annotated teeth per image is 36.78 in training and 37.34 in validation.

**Figure 1: Example panoramic radiograph from the clinical dataset used in this study.**

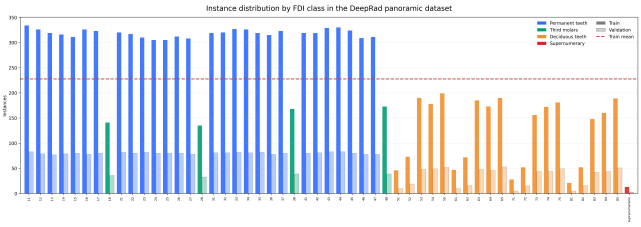



**Figure 2: Distribution of the 53 classes in the panoramic dataset. The validation split preserves full class coverage despite strong imbalance in rare categories.**

## 4.2 Dental Numbering System and Objectives

All labels follow the FDI two-digit notation. Therefore, a correct prediction must satisfy two conditions simultaneously: it must isolate an individual tooth instance and it must assign the correct class identifier. Tooth counting is then derived from the number of predicted instances per image, while tooth numbering is obtained from the predicted FDI classes associated with those instances.

## 4.3 Data Annotation Procedure

The annotations were stored as polygonal instance masks in YOLO segmentation format. Each line encodes a class identifier followed by a sequence of normalized polygon vertices. Across the full dataset, polygons ranged from 5 to 229 vertices, which is adequate to represent irregular crown and root contours without collapsing the problem to rectangular boxes.

## 4.4 Image Preprocessing and Data Split Strategy

All experiments used the project split of 328 training and 82 validation images. Images were resized to 640 pixels on the long side according to the Ultralytics training configuration. For the baseline experiments, we kept the default YOLO training setup for each scale and relied on early stopping where applicable. The augmented YOLOv8m recipe introduced moderate HSV perturbation, translation, scaling, horizontal flip, random erasing, and additional non-spatial Albumentations transforms such as CLAHE, brightness/contrast adjustment, gamma perturbation, and mild image compression.

**Table 2: Experimental configurations used in this study.**

| Configuration | Base model | Batch | Epochs |
|---|---|---|---|
| Baseline N | YOLOv8n-seg | 16 | 100 |
| Baseline S | YOLOv8s-seg | 12 | 50 |
| Baseline M | YOLOv8m-seg | 8 | 57 |
| Optimized M | YOLOv8m-seg | 8 | 56 |
| Post-processed M | YOLOv8m-seg | – | – |

## 4.5 Model Configurations

Table 2 summarizes the experimental configurations used in this paper. The post-processing row reuses the weights from the optimized YOLOv8m model and only changes the inference operating point.

The original article compared Mask R-CNN and YOLO-based segmentation at the conceptual level. In this extended manuscript, the reproducible reruns and added validation analyses focus on the YOLOv8 family, while Mask R-CNN is retained as the original reference architecture in the framing of the study and in the explicit qualitative comparison of Figure 4.

## 4.6 Extended Evaluation Protocol

We report mask precision, mask recall, F1-score, mAP50, and mAP50–95 for the core comparison. To extend the original evaluation, we also report group-level performance, five-fold robustness, and tooth-count metrics computed from the validation audit: mean absolute error (MAE) in tooth counts, exact count agreement, agreement within ±1 and ±2 teeth, and dentition-type agreement.

To assess robustness beyond the fixed split, we performed five-fold cross-validation with the YOLOv8m recipe under the same 640-pixel setup. Each fold contained 328 training images and 82 validation images, with no external data mixed into validation.

# 5 EXPERIMENTAL RESULTS

## 5.1 Training Environment

The experiments were executed with PyTorch 2.9.1 and Ultralytics 8.4.21 on a workstation equipped with a single NVIDIA GeForce RTX 5060 Ti GPU with approximately 8 GB of VRAM. This setup is representative of an accessible single-GPU research environment rather than a multi-GPU server.

## 5.2 Comparative Results on the Fixed Validation Split

Table 3 reports the main comparative results. YOLOv8s-seg achieved the highest mAP50 (0.9003), whereas YOLOv8m-seg achieved the highest mAP50–95 (0.6673). The optimized medium recipe and post-processing remained competitive, but neither surpassed the plain medium baseline on the stricter ranking metrics.

## 5.3 Extended Validation: Clinical Group Analysis

Performance is not uniform across clinical groups. Table 4 shows that permanent teeth dominate the overall score, while deciduous

**Table 3: Performance of the compared configurations on the fixed validation split. The post-processing row uses the optimized YOLOv8m weights with 'conf=0.25', 'iou=0.50', and 'min_area_ratio=0.001'.**

| Model | Precision | Recall | F1 | mAP50 (seg) | mAP50–95 (seg) |
|---|---|---|---|---|---|
| YOLOv8n-seg | 0.7478 | 0.8159 | 0.7804 | 0.8142 | 0.5871 |
| YOLOv8s-seg | 0.8604 | 0.8983 | 0.8789 | **0.9003** | 0.6544 |
| YOLOv8m-seg | **0.8632** | 0.8791 | 0.8711 | 0.8974 | **0.6673** |
| YOLOv8m-seg + expanded augmentation | 0.8403 | 0.8737 | 0.8567 | 0.8807 | 0.6447 |
| YOLOv8m-seg + augmentation + post-processing | 0.8216 | 0.8750 | 0.8474 | 0.8721 | 0.6660 |

**Table 4: mAP50 (seg) by clinical group for representative configurations.**

| Group | YOLOv8s | YOLOv8m | M + PP |
|---|---|---|---|
| Deciduous | 0.8790 | 0.8872 | 0.8386 |
| Permanent (no 3rd molars) | 0.9480 | 0.9394 | 0.9225 |
| Third molars | 0.8974 | 0.8789 | 0.9052 |
| Supranumerary | 0.0000 | 0.0000 | 0.0000 |

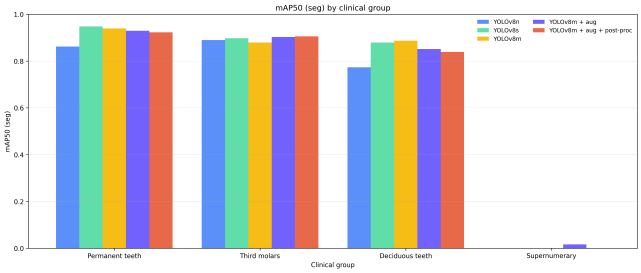



**Figure 3: Group-level comparison across the full experimental set. Permanent teeth are the most stable group, whereas deciduous teeth and the supranumerary class remain the main bottlenecks.**

**Table 5: Five-fold summary for the YOLOv8m protocol.**

| Metric | Mean ± std | Best fold | Worst fold |
|---|---|---|---|
| mAP50 (seg) | 0.8770 ± 0.0132 | 0.9007 | 0.8598 |
| mAP50–95 (seg) | 0.6374 ± 0.0165 | 0.6661 | 0.6163 |
| Precision | 0.8388 ± 0.0208 | 0.8704 | 0.8155 |
| Recall | 0.8610 ± 0.0162 | 0.8843 | 0.8381 |
| F1 | 0.8496 ± 0.0145 | 0.8773 | 0.8360 |

teeth are consistently harder. Third molars remain competitive under the YOLOv8m post-processing setting, but the supranumerary class is still unresolved in all compared models.

## 5.4 Extended Validation: Five-Fold Robustness

To test stability beyond the fixed split, we trained five YOLOv8m folds. Table 5 summarizes the aggregate result. The mean mAP50 was 0.8770 ± 0.0132 and the mean mAP50–95 was 0.6374 ± 0.0165, with the best fold reaching 0.9007 mAP50. These values indicate moderate variance but no catastrophic instability.

**Table 6: Counting audit for the official post-processed pipeline on validation.**

| Audit metric | Value |
|---|---|
| Mean absolute count error | 2.49 teeth/image |
| Exact count agreement | 19 / 82 (23.2%) |
| Agreement within ±1 tooth | 42 / 82 (51.2%) |
| Agreement within ±2 teeth | 48 / 82 (58.5%) |
| Dentition-type agreement | 81 / 82 (98.8%) |

## 5.5 Supplementary Analysis of the Post-processed Medium Configuration

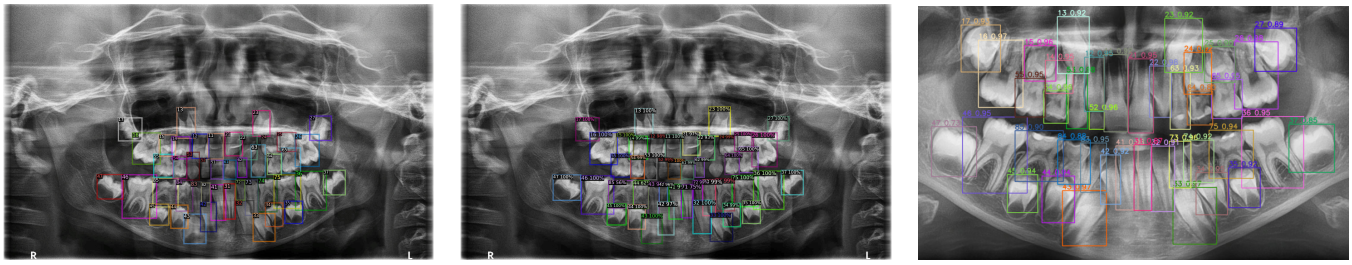
To complement the comparative results, we audited the post-processed medium configuration at the official project operating point. Table 6 shows that the mean absolute counting error was 2.49 teeth per image. Exact tooth-count agreement was obtained in 19 of 82 validation images, while 42 images were within ±1 tooth and 48 were within ±2 teeth. Dentition-type agreement was achieved in 81 of 82 validation cases, and no false mixed-dentition predictions were observed in the audit. Figure 4 revisits the original comparative perspective with explicit numbered views of the ground truth, the Mask R-CNN reference, and the post-processed YOLO prediction.

## 5.6 Discussion of Results

Three main comparative patterns emerge from the experiments. First, the strongest fixed-split result did not come from the largest recipe. YOLOv8s-seg achieved the best mAP50, while YOLOv8m-seg achieved the best mAP50–95. This suggests that the medium configuration benefits ranking quality under stricter overlap thresholds, but the simpler small configuration remains highly competitive and may be the better default when computational budget matters.

Second, stronger training interventions were not automatically beneficial. The optimized YOLOv8m recipe with expanded augmentation underperformed the plain YOLOv8m baseline on both mAP50 and mAP50–95. This result is consistent with the notion that panoramic radiographs have strong geometric regularity, so overly aggressive augmentation may disturb clinically meaningful alignment rather than improve generalization.

Third, the supplementary post-processing stage is clinically useful but not a universal remedy. Relative to the same optimized model at the default Phase 4 operating point, calibrated NMS and the area threshold produced a small F1 gain (0.8468 to 0.8474) and a marginal mAP50–95 gain (0.6659 to 0.6660). However, these



(a) Ground truth with FDI numbering.

(b) Mask R-CNN reference prediction.

(c) Post-processed medium-configuration prediction with counted and numbered teeth.

**Figure 4: Explicit counting and numbering examples. The original comparative perspective is preserved through the side-by-side visualization of ground truth, Mask R-CNN, and YOLOv8 on a representative case, while the numbered YOLO overlay remains aligned with the extended evaluation reported in this revision.**

improvements were not enough to surpass the simpler baseline-medium model. The main value of post-processing is therefore not leaderboard improvement, but better control of anatomically implausible predictions during case review.

The group analysis and the counting audit converge on the same conclusion: permanent teeth are substantially easier than mixed-dentition and rare classes. Exact count agreement was below one quarter of the validation set, but dentition-type agreement was nearly perfect. This means the model is already reliable for coarse structural interpretation, yet still not precise enough to replace manual review when exact tooth count is clinically critical.

## 6 THREATS TO VALIDITY

**Internal validity.** All experiments were conducted on a single institutional dataset. Although the train/validation split preserves all 53 classes in validation, the study does not include an external test cohort. The reported values may therefore retain acquisition-specific biases tied to a single scanning workflow.

**Construct validity.** The reported metrics are segmentation-oriented. Clinical use, however, depends on correct tooth counts and FDI numbering. The dedicated counting audit narrows this gap, but clinically grounded and workflow-oriented error measures remain necessary.

**Conclusion validity.** The five-fold study reduces dependence on a single split, but the counting audit was performed on the official validation operating point of the supplementary post-processed medium configuration. Further comparison of count error across all variants would strengthen the argument.

**External validity.** The dataset is private, single-center, and clinically curated. Generalization to other scanners, age distributions, and acquisition protocols remains open. The rare supranumerary class is especially underrepresented, limiting conclusions for uncommon dental conditions.

## 7 CONCLUSION AND FUTURE WORK

This paper preserves the original comparative objective of evaluating Mask R-CNN and YOLO-based segmentation for automatic tooth counting and FDI numbering in panoramic radiographs, while extending the validation protocol with group-level analysis, five-fold robustness, and a dedicated counting audit. On the fixed validation split, YOLOv8s-seg achieved the best mAP50 (0.9003), while YOLOv8m-seg achieved the best mAP50–95 (0.6673). The five-fold study showed stable performance, with mean scores of $0.8770 \pm 0.0132$ for mAP50 and $0.6374 \pm 0.0165$ for mAP50–95. The counting audit showed that the post-processed medium configuration preserved dentition-type coherence in 98.8% of cases, although exact counts remain challenging in mixed-dentition images.

The extended evidence strengthens the comparative conclusion that the YOLOv8 small and medium configurations define the strongest practical trade-off within the reproducible branch reported here, while Mask R-CNN remains part of the original comparative framing of the work. The post-processed medium configuration is useful for supplementary qualitative analysis and case review, but it does not change the overall ranking of the compared YOLO models. Future work should prioritize external validation, broader inter-architecture comparison, more support for rare dental classes, and anatomically informed reasoning over the FDI sequence. Future work will re-establish a fully symmetric rerun of both architectures under the extended protocol.

## ACKNOWLEDGMENT

The authors thank Samuel Starke for making the Kaggle release of the dataset publicly available.

## REFERENCES

[1] Burak Dayi, Huseyin Uzen, Ipek Balikci Cicek, and Suayip Burak Duman. 2023. A Novel Deep Learning-Based Approach for Segmentation of Different Type Caries Lesions on Panoramic Radiographs. *Diagnostics* 13, 2 (2023), 202. https://doi.org/10.3390/diagnostics13020202

[2] DENTEX Challenge Organizers. 2023. DENTEX: An Abnormal Tooth Detection with Dental Enumeration and Segmentation Challenge. In *MICCAI Challenge Proceedings*. Grand Challenge on Dental Enumeration and Diagnosis.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969. https://doi.org/10.1109/ICCV.2017.322

[4] Ya-Yun Huang, Chiung-An Chen, Yi-Cheng Mao, Chih-Han Li, Bo-Wei Li, Tsung-Yi Chen, Wei-Chen Tu, and Patricia Angela R. Abu. 2025. An Integrated System for Detecting and Numbering Permanent and Deciduous Teeth Across Multiple Types of Dental X-Ray Images Based on YOLOv8. *Diagnostics* 15, 13 (2025), 1693. https://doi.org/10.3390/diagnostics15131693

[5] Robert P. Langlais, Olaf E. Langland, and Carel J. Nortje. 2020. *Diagnostic Imaging of the Jaws* (2nd ed.). Wiley-Blackwell.

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE*

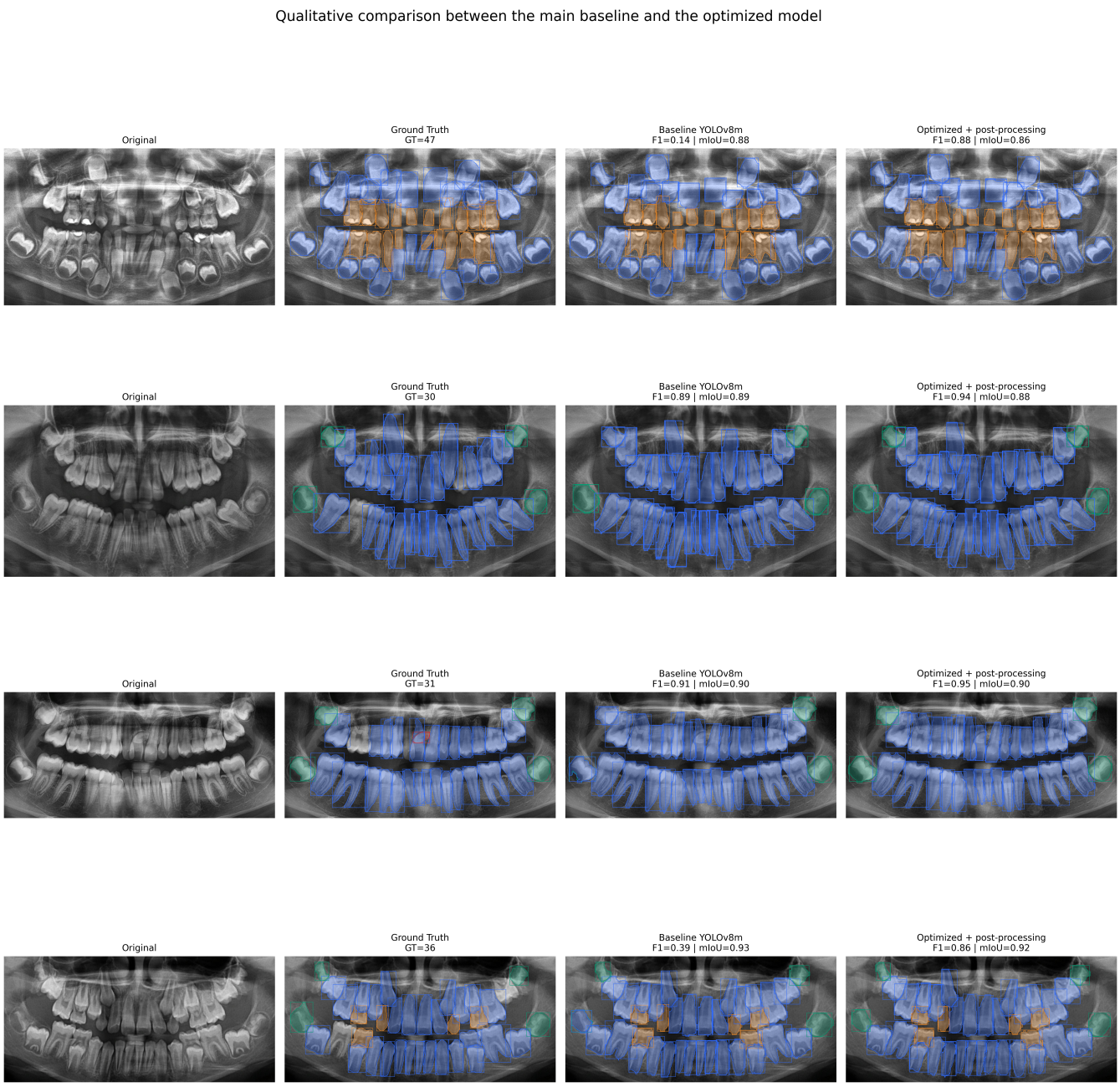



**Figure 5: Qualitative comparison between representative predictions. The panels contrast cleaner permanent cases with more difficult mixed-dentition examples, showing that the strongest compared configurations remain robust on well-separated permanent teeth but degrade in crowded eruption patterns.**

*Conference on Computer Vision and Pattern Recognition*. 779–788. https://doi.org/10.1109/CVPR.2016.91

[7] A. T. Rodrigues, M. G. Silva, and F. L. Costa. 2022. Artificial intelligence in dental radiography: Current applications and future perspectives. *Computers in Biology and Medicine* 140 (2022), 105042. https://doi.org/10.1016/j.compbiomed.2021.105042

[8] Rata Rokhshad, Maxime Ducret, Akhilanand Chaurasia, Teodora Karteva, Miroslav Radenkovic, Jelena Roganovic, Manal Hamdan, Hossein Mohammad-Rahimi, Joachim Krois, Pierre Lahoud, and Falk Schwendicke. 2023. Ethical considerations on artificial intelligence in dentistry: A framework and checklist. *Journal of Dentistry* 135 (2023), 104593. https://doi.org/10.1016/j.jdent.2023.104593

[9] Falk Schwendicke, Wojciech Samek, and Joachim Krois. 2020. Artificial intelligence in dentistry: Chances and challenges. *Journal of Dental Research* 99, 7 (2020), 769–774. https://doi.org/10.1177/0022034520915714

[10] Samuel Starke. 2025. DeepRAD Panoramic Dental X-Rays YOLOv8 Numbering. Kaggle dataset page. Kaggle dataset, version 1. Modified: 2025-05-16. Accessed: 2026-03-08.

[11] Ultralytics. 2023. YOLOv8 Documentation. https://docs.ultralytics.com. Accessed: 2026-03-08.

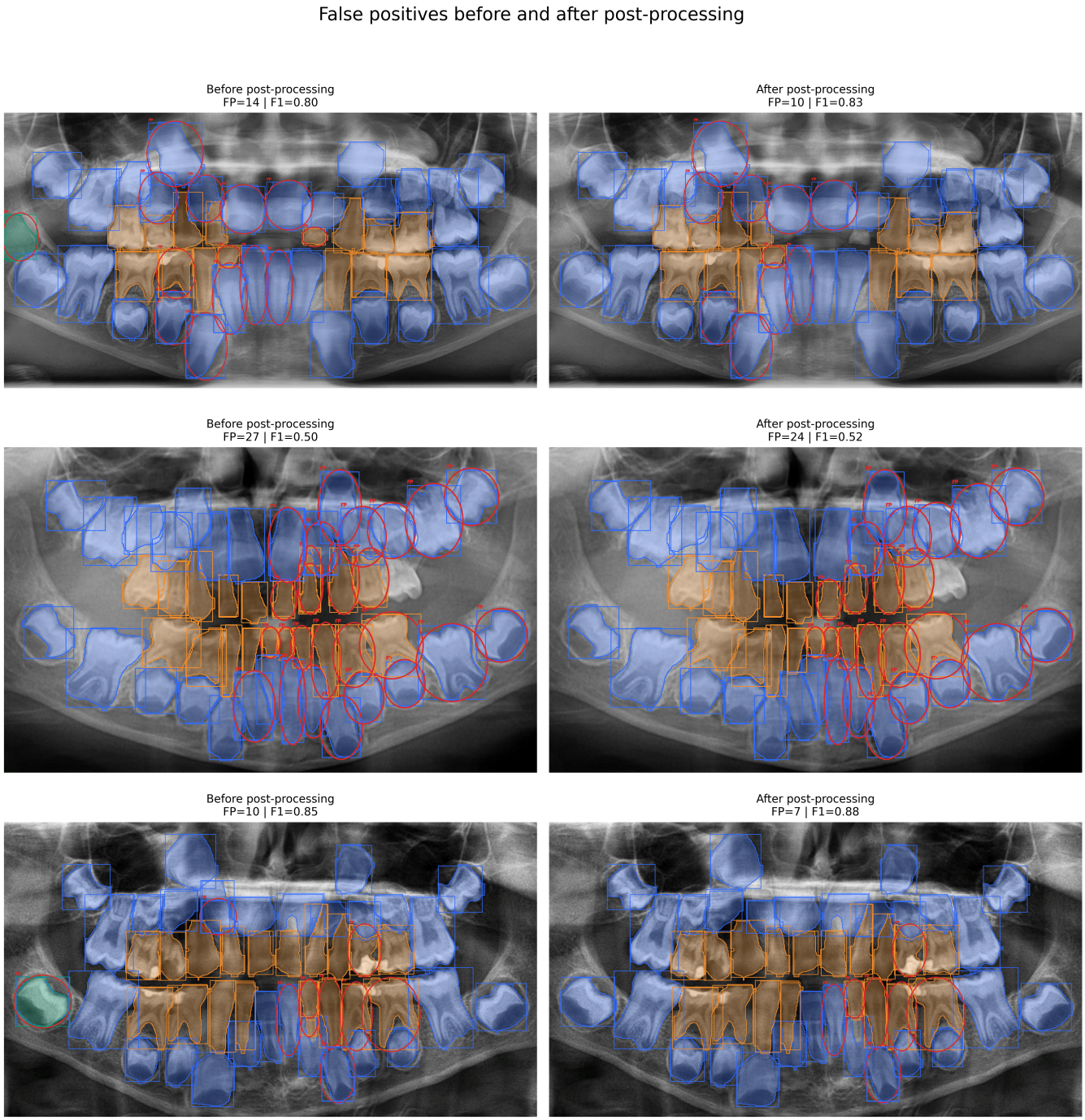



**Figure 6: False-positive analysis before and after post-processing. The calibrated operating point reduces anatomically implausible detections in crowded regions, but it does not eliminate all counting errors.**

[12] Ann Wenzel. 2021. Radiographic modalities for diagnosis of caries in a historical perspective: from film to machine-intelligence supported systems. *Dentomaxillofacial Radiology* 50, 5 (2021). https://doi.org/10.1259/dmfr.20210010

[13] Stuart C. White and Michael J. Pharoah. 2018. *Oral Radiology: Principles and Interpretation* (7th ed.). Elsevier.