Anatomy-Constrained Duplication Loss for Tooth Detection in Panoramic Radiographs

Noam Delbari[1,2] · **Ariel Hirschhorn**[3] · **Bella Specktor-Fadida**[2]

[1]School of Computer Science, Reichman University, Herzliya, Israel · [2]Department of Medical Imaging Sciences, University of Haifa, Haifa, Israel · [3]Department of Cranio-Maxillofacial Surgery, Sheba Medical Center, Ramat Gan, Israel

Healthcare Computer Vision (hVision) Lab, University of Haifa

ODIN 2026
Oral & Dental Image Analysis Workshop
MICCAI 2026 · Strasbourg
Paper #28

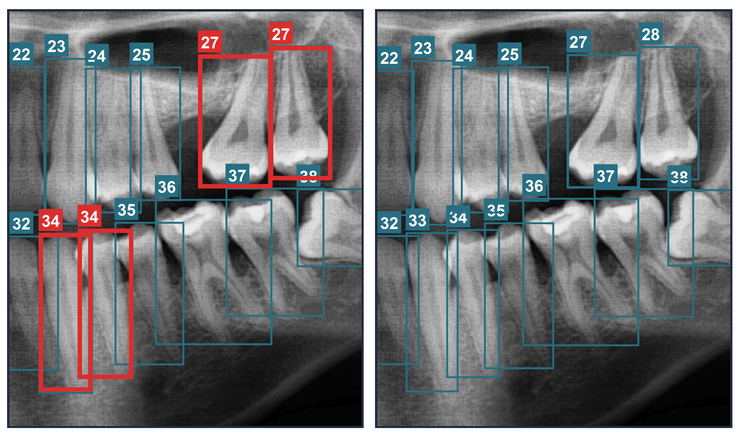
Motivation

Tooth detection is routinely formulated as 32-class instance detection over the FDI two-digit nomenclature [1]. The classification objective is evaluated independently for each candidate box, but the anatomical constraint that a given tooth number occurs at most once per dentition is never expressed during training.

1.5–2.4
duplicated tooth numbers per image — unmodified YOLOv11n [3] at 1024 px

5–7
duplicated tooth numbers per image — unmodified YOLOv11n [3] at 640 px

Each duplication corrupts the dental chart derived from the detection and propagates into every per-tooth diagnostic and treatment-planning step built upon it [2].



Baseline — three numbers predicted twice: **27, 34, 46**

+ $\mathscr{L}_{\text{dup}}$ — no number predicted twice

Fig. 1. ■ number predicted more than once ■ unique number. Identical image and seed (HITL validation image 294, seed 789, 1024 px). Against 31 annotated teeth the baseline emits 34 boxes, the constrained model exactly 31.

Related work

NMS, Soft-NMS [5] — overlap-based; duplicated numbers sit on distinct teeth (median IoU 0.13 / 0.23), so ≤ 19% of duplicate pairs are removed.

Per-class deduplication — reaches zero duplications, but discards genuine teeth (recall −1.0 to −1.6) and returns no learning signal.

Set prediction, DETR [6] — one-to-one matching by architecture, at the cost of a different detector family and slower inference.

Numbering networks [2,7] — report enumeration accuracy, but delegate uniqueness to post-processing rules.

Our contribution. A soft, differentiable duplication loss penalising any tooth number whose aggregated probability exceeds one — no extra parameters, no architectural change, no inference overhead.

Method

Penalise probabilities summing above one. For every tooth number, the class probabilities of the selected representatives are aggregated into a soft count, and any probability mass beyond unity is charged as a hinge penalty. Each stage below carries the expression it computes.

Detector head — class logits

$s \in \mathbb{R}^{A \times 32}$

YOLOv11n [3]: A anchors scored over 32 FDI classes, trained with the unmodified BCE objective.

↓

① One representative per annotated tooth

$a^*_n = \text{argmax}_{a \in A_n} s_{\text{align}}(a, c_n)$,

$s_{\text{align}} = {s_{\text{cls}}}^{0.5} \cdot \text{IoU}^6$

a_1 0.42	**a_2 0.71 ✓**	a_3 0.35

Of the anchors task-aligned assignment [4] gives tooth 14, the highest-scoring one is retained; N teeth yield N representatives.

↓

② Row-softmax; column sums are soft counts

$p_{ij} = e^{s_{ij}} / \Sigma_k e^{s_{ik}}$, $q_j = \Sigma_{i=1}^{N} p_{ij}$

rep.	13	14	21	22	…
1	.90	.08	.01	.01	…
2	.06	**.62**	.30	.02	…
3	.02	**.57**	.40	.01	…
4	.01	.00	.29	.70	…
q_j	**0.99**	**1.27**	**1.00**	**0.74**	…

Rows are distributions over the 32 numbers. Number 14 is claimed by two representatives (q = 1.27); number 21 is exactly saturated.

↓

③ Hinge penalty on the excess mass

$\mathscr{L}_{\text{dup}} = \Sigma_{j=1}^{C} \max(0, q_j - 1) = 0.27$ (example above)

$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{standard}} + \lambda_d \mathscr{L}_{\text{dup}}, \lambda_d = 0.5$

gradient to class logits only

Fig. 2. The duplication term acts entirely inside the training loop. Its gradient, $\partial \mathscr{L}_{dup} / \partial s_{ik} = p_{ik}(\mathbb{1}[k \in A] - m_i)$ with $A = \{j : q_j > 1\}$, reaches the classification logits alone (dashed). Inference is untouched.

An upper bound, not an equality. A number left unclaimed ($q_j < 1$) is never penalised, and a per-class budget $u_j = 2$ accommodates a documented supernumerary tooth.

Redistribution rather than suppression. The gradient sums to zero across an anchor's classes: probability mass is reassigned between numbers, objectness is left untouched, and the term vanishes once no number exceeds its budget.

The softmax is confined to the loss. The BCE classification head and the sigmoid confidences used at inference remain unchanged, and retaining a single representative per tooth preserves the positive redundancy of task-aligned assignment.

Experimental results

Datasets

	HITL 2023	DENTEX 2023 [7]
Images (train / val)	595 (506 / 89)	634 (538 / 96)
Labels	instance masks, 32 FDI	boxes, 32 FDI
Detector	YOLOv11n-seg	YOLOv11n

HITL is drawn from Kaggle (Humans in the Loop) and DENTEX from the MICCAI 2023 challenge across three hospitals. Both detectors are COCO-initialised and trained at 1024^2 with AdamW for at most 100 epochs under early stopping, over five seeds (42, 123, 456, 789, 2024); λ_d = 0.5 is fixed a priori and significance is assessed by paired t-tests. All metrics are evaluated at a fixed confidence threshold of 0.25 and compared within a dataset only.

Detection accuracy and duplication

Means over five seeds. **Bold** denotes the best value per metric within a resolution block. *p<0.05, **p<0.01 against the baseline at the same resolution (paired t-test, n = 5). Reference detectors (DETR [6], Mask R-CNN, Faster R-CNN) operate on short-side-800 input and are reported for context only.

Res	Model	mAP50	P	R	Dups ↓
HITL 2023 (89 validation images)					
640	Baseline	0.868	0.746	0.854	4.95
640	+ $\mathscr{L}_{\text{dup}}$	**0.894****	**0.776***	**0.866**	**3.68****
1024	Baseline	0.949	0.869	0.936	1.53
1024	+ $\mathscr{L}_{\text{dup}}$	**0.951**	**0.874**	**0.940**	**1.31**
ref	DETR R50	0.919	0.801	0.947	2.63
ref	Mask R-CNN R50	0.944	0.789	0.829	3.20
DENTEX 2023 (96 validation images)					
640	Baseline	0.844	0.707	**0.858**	7.01
640	+ $\mathscr{L}_{\text{dup}}$	**0.862***	**0.737**	0.851	**5.76**
1024	Baseline	**0.952**	0.861	0.952	2.43
1024	+ $\mathscr{L}_{\text{dup}}$	0.951	**0.873****	**0.955**	**1.92***
ref	DETR R50	0.847	0.851	0.902	0.38
ref	Faster R-CNN R50	0.913	0.568	0.427	1.45

Duplications decrease and precision increases in every dataset–resolution cell; recall declines by at most 0.7 points and improves in three of the four. DETR reaches 0.38 duplications on DENTEX only at a deficit of ten mAP50 points.

Across inference resolutions

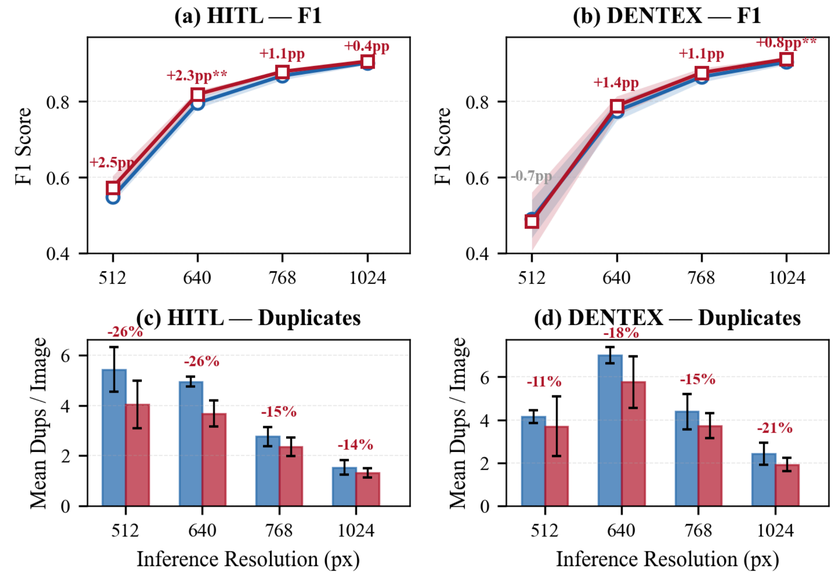


Fig. 3. Inference at 512–1024 px, mean ± SD over five seeds. **Top:** F1 gain in points. **Bottom:** duplications per image and relative reduction. Duplications decrease in all eight cells and F1 improves in seven; at n = 5 only the largest effects attain p < 0.05.

Training-time loss vs. inference-time fixes

At 1024 px under a common evaluator, means over five seeds. Dedup retains the most confident box per FDI number, which is equivalent to own-class Hungarian matching.

Method	HITL			DENTEX		
	F1	R	Dups	F1	R	Dups
Baseline	0.902	0.937	1.53	0.904	0.952	2.43
+ Soft-NMS	0.906	0.936	1.20	0.913	0.950	1.76
+ tuned class-aware NMS	0.906	0.936	1.22	0.911	0.951	1.92
+ Dedup	0.919	0.927	**0.00**	0.926	0.936	**0.00**
+ $\mathscr{L}_{\text{dup}}$	0.906	**0.940**	1.32	0.912	**0.955**	1.93
+ $\mathscr{L}_{\text{dup}}$ + Dedup	**0.921**	0.932	**0.00**	**0.932**	0.944	**0.00**

Deduplication attains zero duplications by sacrificing recall ($p \leq 0.001$), whereas the loss maintains recall at or above the baseline; in combination they give the highest F1 on both datasets. NMS variants leave most duplications in place, since same-number boxes rarely overlap.

Conclusions

- A single differentiable anatomical constraint reduces duplicated tooth numbers by 14–26% and improves precision at every resolution examined, at no inference cost.
- The loss is complementary to per-class deduplication: in combination, F1 attains 0.921 (HITL) and 0.932 (DENTEX), the highest value on either dataset.
- Gains are largest where the detector is weakest (640 px: +2.6 mAP50, +2.3 F1, −26% duplications on HITL); effect size remains dependent on dataset and operating point.

Limitations. The loss constitutes an upper bound only: it penalises neither missing nor phantom teeth, does not adjudicate which tooth owns a contested number, and aggregate counts of incorrect numbers do not establish individual relabelling. Both datasets contain permanent dentition only, the evidence is confined to YOLOv11n, and a residual AutoBatch batch-size confound cannot be excluded.

References

1. Fédération Dentaire Internationale. Two-digit system of designating teeth. *Int Dent J* 21, 104–106 (1971).
2. Tuzoff DV et al. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofac Radiol* 48, 20180051 (2019).
3. Jocher G, Qiu J. Ultralytics YOLO11 (2024).
4. Feng C et al. TOOD: Task-aligned one-stage object detection. *ICCV* (2021).
5. Bodla N et al. Soft-NMS: improving object detection with one line of code. *ICCV* (2017).
6. Carion N et al. End-to-end object detection with transformers. *ECCV* (2020).
7. Hamamci IE et al. DENTEX: an abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic X-rays. arXiv:2305.19112 (2023).

Partially supported by the Israeli Ministry of Innovation, Science and Technology and the Planning and Budgeting Committee of the Council for Higher Education.

Contact: bspecktor@univ.haifa.ac.il