Hybrid Frequency–Spatial Attention and Arch-Aware Priors for Tooth Detection and FDI Numbering in Dental Images

Xi Wu*, Ruijie Huang†✉, Jiangping Zhu*‡✉, Shiquan Min*, Pei Zhou*, Zhongjian Wang§
*College of Computer Science, Sichuan University, Chengdu 610065, China
†West China School/Hospital of Stomatology, Sichuan University, Chengdu 610065, China
‡School of Information Science and Technology, Tibet University, Lhasa 850000, China
§MoEntropy Science (Chengdu) Pharmaceutical Technology Co., Ltd., Chengdu 610094, China
✉ Corresponding authors: ruijmhuang@gmail.com; zjp16@scu.edu.cn

***Abstract*—Current oral imaging modalities, such as panoramic radiographs and intraoral photographs, suffer from challenges including noise, blurring, occlusions, and significant variations in pose and scale. Intraoral photographs further exhibit specular reflections and complex illumination. Moreover, tooth arrangement is governed by strict morphological and anatomical constraints. These factors collectively pose significant challenges to accurate tooth position detection. To address these challenges, we propose a novel network architecture. Specifically, we introduce a Hybrid Frequency–Spatial Attention (HFSA) capable of simultaneously capturing both frequency and spatial features from images, thereby enhancing the network's robustness against various noise conditions. Furthermore, we design a Differentiable Class-Uniqueness Gating (UniGate) Mechanism to impose unique indexing constraints on each tooth. Additionally, we develop an Arch-aware Query Graph Prior (QGP), which enforces morphological and anatomical constraints on the detection process. We validate the performance of our approach through experiments on a public DSLR intraoral dataset, a private smartphone in-the-wild intraoral photograph dataset, and the public DENTEX2023 benchmark. Our method achieves state-of-the-art performance across all three datasets with AP scores of 78.3%, 54.4%, and 54.1%, respectively.**

***Index Terms*—Tooth detection, FDI numbering, Dental imaging, Deep learning, Transformers, Hybrid frequency-spatial attention, Graph-based priors.**

I. Introduction

Automated tooth numbering technology can assign FDI tooth notation [1] to teeth based on dental imagery, facilitating efficient clinical charting and the digitization of Electronic Dental Records (EDR). However, intraoral photographs frequently suffer from severe artifacts such as specular reflections and motion blur. These artifacts degrade critical visual cues, thereby making it difficult to maintain the strict anatomical consistency mandated by the FDI numbering schema.

Early work adapted Faster R-CNN, SSD, and YOLO [2]–[4] to dental radiographs and related dental imaging tasks [5]–[8]. These pipelines deliver high throughput but still treat tooth numbering as independent per-box classification. Anchor designs and non-maximum suppression (NMS) require careful tuning, and rule-based post-processing is often introduced to remove duplicates and repair inconsistent FDI labels. Under occlusion, crowding, label noise, and cross-device domain shift, such heuristics frequently fail and can yield clinically invalid numberings even if the bounding boxes are roughly correct.

Transformer-based detectors such as DETR [9] cast detection as set prediction with bipartite matching and eliminate anchors and NMS. Deformable DETR [10] adds sparse multi-scale attention for small objects. Subsequent variants refine query design [11], [12], introduce denoising-based training [13], [14], and explore one-to-many / hybrid matching strategies to accelerate convergence [15], [16]. Recent models such as DEIM [17] and D-FINE [18] further improve matching and bounding-box regression and form strong baselines on natural images. Despite their extensive application in medical imaging, these category-agnostic models inherently overlook the global FDI uniqueness and dental arch topology essential for clinical consistency. Furthermore, the significant domain gap between natural scene and dental imagery characterized by noise, blur, and low contrast hinders their direct applicability.

Within the dental domain, teeth possess predictable geometry, quadrant symmetry, and a fixed FDI sequence. Leveraging these priors, previous methods [19]–[21] employ adjacency graph modules or uniqueness constraints to restrict each tooth identity to a single detection. However, the reliance of these approaches on heavy relational branches or heuristic post-processing complicates deployment and preserves sensitivity to intermediate errors.

To address these challenges, we propose a novel, prior-aware detection framework. Specifically, we introduce a Hybrid Frequency-Spatial Attention (HFSA) module capable of simultaneously capturing both frequency and spatial features, thereby enhancing network robustness against complex noise and scale variations. To ensure clinical consistency, we introduce a Differentiable Class-Uniqueness Gating (UniGate) mechanism to enforce one-to-one FDI indexing. Combined with a Query Graph Prior (QGP), these components embed essential morphological and spatial topologies directly into the detection pipeline. Validation across a public DSLR dataset, a private smartphone in-the-wild intraoral-photograph dataset,

and the public DENTEX2023 benchmark confirms the effectiveness of these contributions.

- We propose the HFSA module, which fuses spectral analysis with spatial attention to explicitly decouple structural features from noise, thereby significantly enhancing robustness in uncontrolled imaging scenarios.
- We introduce UniGate and QGP to inject strict FDI logic into the detection pipeline. UniGate reduces duplicated FDI identities via differentiable gating on FDI-indexed datasets, while QGP enforces morphological and topological consistency via graph-based priors, eliminating the need for heuristic post-processing.
- We demonstrate the superior generalizability of our method through extensive experiments on three diverse datasets. Our method consistently achieves state-of-the-art detection AP across all datasets.

II. Proposed Method

Overview. Overview. Fig. 1 illustrates our architecture. We employ a CNN-Transformer hybrid encoder enhanced by HFSA modules for blur-robust feature extraction. The decoder utilizes a deformable transformer with a multi-inquiry scheme for query initialization and refinement. To leverage domain knowledge, we propose two priors: UniGate, which prevents duplicate tooth indexing, and QGP, which constrains query interactions based on dental arch topology. For datasets annotated with tooth categories, UniGate is disabled because multiple instances can share the same category, while QGP remains applicable as a topology prior.

A. HFSA

We extract multi-scale features and build a feature pyramid, with all levels projected to a shared dimension D. HFSA is applied to encoder-refined features and pyramid outputs to enhance global context and multi-scale details.

Given an input feature $x \in \mathbb{R}^{D\times H\times W}$, HFSA contains two parallel paths and a gated residual fusion. The frequency path emphasizes structural details that are often weakened by blur and highlights. We compute the real FFT $\hat{x} = \mathcal{F}(x)$ and construct K concentric low-pass masks $\{M_k\}_{k=1}^{K}$ with increasing cutoff radii. Let $x_{\text{lp}}^{(k)} = \mathcal{F}^{-1}(M_k \odot \hat{x})$. We form a band decomposition using residual differences:

$$\begin{aligned} x_{\text{band}}^{(1)} &= x_{\text{lp}}^{(1)}, \\ x_{\text{band}}^{(k)} &= x_{\text{lp}}^{(k)} - x_{\text{lp}}^{(k-1)}, \qquad k = 2, \dots, K, \\ x_{\text{band}}^{(K+1)} &= x - x_{\text{lp}}^{(K)}. \end{aligned} \tag{1}$$

Each band is modulated by a lightweight depthwise/grouped convolution producing a spatial weight map $w_k(x)$, and the frequency-enhanced feature is

$$x^{\text{freq}} = \sum_{k=1}^{K+1} w_k(x) \odot x_{\text{band}}^{(k)}. \tag{2}$$

This design preserves smooth structures via $x_{\text{lp}}^{(K)}$ while selectively amplifying edge-like and texture-like signals from the residual terms.

In parallel, the spatial path injects an inductive bias aligned with dental morphology. We use depthwise strip convolutions to capture long-range horizontal context along the dental arch and vertical elongated tooth patterns:

$$x^{\text{strip}} = \text{Conv}_{1\times 1}\Big(\text{DWConv}_{1\times k}(x) + \text{DWConv}_{k\times 1}(x)\Big), \tag{3}$$

where a relatively large k (e.g., $k = 11$) covers multiple neighboring teeth while remaining efficient due to depthwise computation.

Finally, HFSA fuses the two paths with a SK-style gate conditioned on their combined response. Let $s = x^{\text{freq}} + x^{\text{strip}}$. We obtain per-channel mixing weights $[\alpha, \beta] = \text{softmax}(\psi(\text{GAP}(s)))$, where $\psi(\cdot)$ is a small 1×1 MLP producing $2D$ scores. The fused feature is $x^{\text{fuse}} = \alpha \odot x^{\text{freq}} + \beta \odot x^{\text{strip}}$, followed by a 1×1 projection and a learnable residual scaling:

$$\tilde{x} = x + \tanh(\gamma) \cdot \text{Conv}_{1\times 1}(x^{\text{fuse}}), \tag{4}$$

where γ is initialized near zero to stabilize optimization. Applying HFSA at multiple scales improves tooth contours and restores fine textures with modest overhead.

B. QGP

Pure self-attention among decoder queries is flexible but may yield anatomically implausible configurations for dentistry, e.g., mixing upper/lower arches or producing inconsistent ordering along the dental curve. We introduce a *QGP* to regularize query interactions by restricting self-attention to a local neighborhood defined on an estimated arch-wise manifold.

QGP is built from the normalized reference points of the *main* queries $\{r_i \in [0,1]^2\}_{i=1}^{Q}$ (excluding denoising queries). Let $r_i = (x_i, y_i)$. We first assign each query to an arch using its vertical coordinate,

$$g_i = \begin{cases} \text{upper}, & y_i < 0.5, \\ \text{lower}, & y_i \geq 0.5, \end{cases} \tag{5}$$

which matches our implementation where only upper/lower grouping is used to avoid breaking connectivity around central incisors. Within each arch, we map the 2D points to a 1D curve parameter t_i that reflects their progression along the dental arch. In practice, we compute t_i by (i) fitting a simple arch curve when the points are sufficiently stable, and (ii) falling back to the horizontal coordinate x_i during early training or when curve fitting is unreliable; this mirrors the robust choice in our implementation.

Given $\{t_i\}$, QGP constructs a local graph by selecting k nearest neighbors on the 1D manifold within the same arch:

$$\mathcal{N}_k(i) = \underset{j:\, g_j = g_i}{\arg\text{topk}}\big(-|t_i - t_j|\big). \tag{6}$$

We then form a boolean attention mask to block all non-neighbor interactions:

$$\tilde{\mathbf{A}}_{ij} = \begin{cases} 0, & j \in \mathcal{N}_k(i) \text{ or } j = i, \\ -\infty, & \text{otherwise}, \end{cases} \tag{7}$$

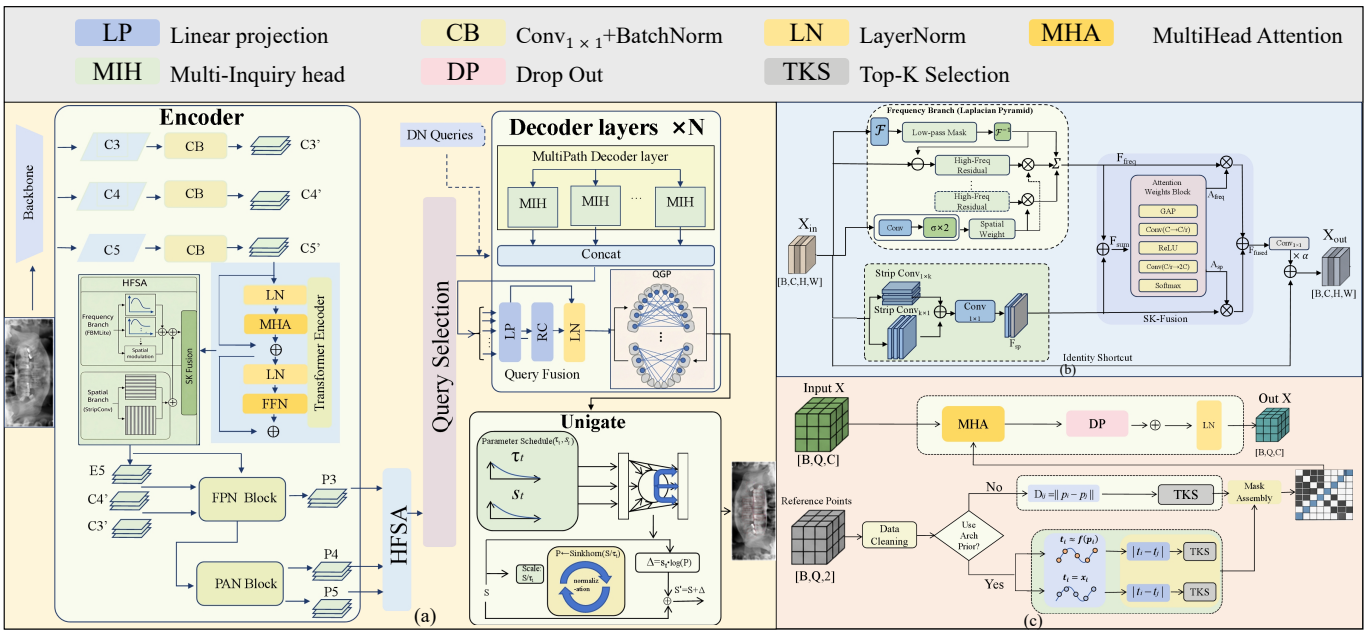


Fig. 1. **Framework of the proposed method.** (a) DETR-style pipeline with an HFSA-enhanced encoder and a deformable decoder equipped with UniGate and QGP. (b) HFSA fuses Fourier-domain context with strip convolutions. (c) QGP builds a dynamic k-NN arch graph to constrain query attention.

and inject it into a dedicated multi-head self-attention layer over the main queries. This effectively turns query self-attention into message passing on a dynamic kNN graph whose edges are recomputed each layer from current reference points.

The graph-updated queries $\hat{q}_i$ are merged back with the original queries using a gated residual update:

$$q_i \leftarrow q_i + \sigma(\lambda)\big(\hat{q}_i - q_i\big), \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid and λ is initialized to a negative value, so QGP starts as a weak prior and gradually becomes effective as training progresses. In our decoder, QGP is applied only on several layers (e.g., intermediate layers) and only to the main queries, leaving denoising queries and cross-attention unchanged. This design suppresses invalid cross-arch interactions and stabilizes arch-consistent layouts, while still allowing local reconfiguration to handle missing teeth, occlusions, and varied intraoral viewpoints.

C. UniGate

In FDI-indexed tooth detection, each anatomical tooth identity should emerge at most once per image. However standard set-prediction losses do not explicitly encode this global uniqueness, so that multiple queries may compete for the same FDI index while other identities remain unused. For this, we introduce UniGate to reshape the classification logits towards a nearly one-to-one assignment between present FDI indices and main queries. UniGate is enabled only on FDI-labeled datasets and disabled on category-labeled detection.

For each training image b, we compute a binary vector $\mathbf{m}_b \in \{0,1\}^C$ to indicate which tooth categories are present at the ground truth. We then extract the logits of the main queries and form a sub-matrix $\mathbf{S}_b \in \mathbb{R}^{Q_m \times C_b}$ over the present classes, where $C_b = \sum_c m_{b,c}$. After transposition and temperature scaling, we conduct a fixed number of row–column normalizations to obtain a doubly stochastic matrix $\mathbf{P}_b \in [0,1]^{C_b \times Q_m}$, which can be viewed as a soft permutation [22] between present categories and queries. Finally, we map it back to the query-by-class layout and add a log-probability penalty to the original logits:

$$\tilde{\mathbf{S}}_b = \mathbf{S}_b + \lambda_t \log \mathbf{P}_b, \tag{9}$$

where λ_t is a training-time strength coefficient. We apply UniGate only to the main queries to avoid interfering with denoising supervision.

D. Loss Function

We follow DEIM training with Hungarian matching between decoder predictions and ground-truth teeth, yielding matched pairs $\mathcal{M}$. For a matched pair $(i,j) \in \mathcal{M}$, the detector predicts a box $\hat{b}_j$ and class logits $\hat{\mathbf{s}}_j$. The detection loss is

$$\mathcal{L}_{\text{det}} = \sum_{(i,j)\in\mathcal{M}} \Big(\lambda_{\text{cls}}\mathcal{L}_{\text{focal}}(\hat{\mathbf{s}}_j, c_i^*) + \lambda_1 \|\hat{b}_j - b_i^*\|_1 + \lambda_{\text{giou}}\mathcal{L}_{\text{GIoU}}(\hat{b}_j, b_i^*)\Big). \tag{10}$$

where c_i^* and b_i^* are the ground-truth FDI label and box.

Overall objective. UniGate reshapes the main-query logits during training and is optimized implicitly through $\mathcal{L}_{\text{det}}$. The total loss is

$$\mathcal{L} = \mathcal{L}_{\text{det}}. \tag{11}$$

We follow DEIM training and apply the same loss on intermediate decoder layers and denoising queries; for brevity, we present only the final-layer form of $\mathcal{L}_{\text{det}}$.

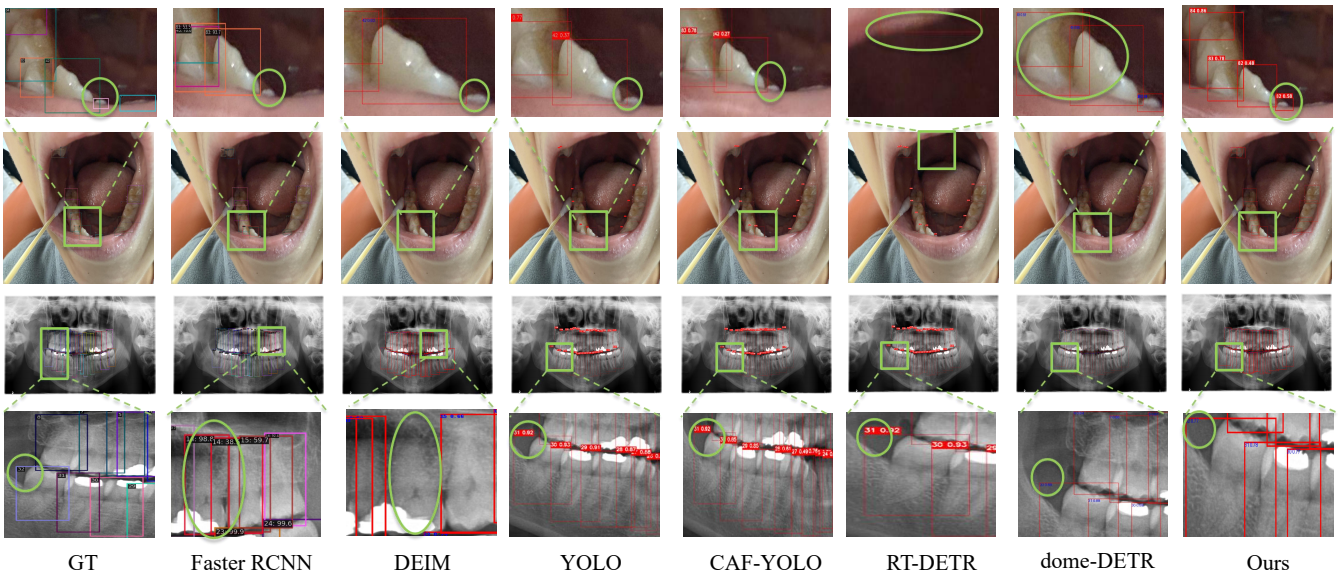


Fig. 2. Visual comparison of detection and numbering results on DENTEX2023 (bottom) and smartphone (top) intraoral photographs.The green boxes and dashed lines indicate the localized enlargements for a detailed comparison of different methods.

III. Experiments

A. Experimental Settings

Datasets and Metric. We evaluate our method on two intraoral-photograph datasets and the DENTEX2023 panoramic-radiograph benchmark [8]. (1) A *DSLR dataset* with **505/112** train/val images and **11,138/2,437** annotated teeth in **7** categories [23]. (2) An in-the-wild *Smartphone dataset* with **878/221** images and **9,837/2,459** instances annotated by **52** FDI indices. (3) The *DENTEX2023 enumeration benchmark* [8] with **570/64** train/val images and **16,234/1,861** instances over **32** FDI classes; test results are obtained via the official server.

For detection evaluation, we report COCO-style bounding-box Average Precision (AP), averaged over IoU thresholds from 0.50 to 0.95 in increments of 0.05. Additionally, we include AP_{50} and AP_{75}, corresponding to AP at IoU = 0.50 and IoU = 0.75, respectively.

Implementation Details.

All models are implemented in PyTorch and optimized with AdamW and a flat–cosine schedule. We train for 160 epochs with a batch size of 16 on NVIDIA 5090D GPUs. Data augmentation follows a Dense O2O-style pipeline with standard geometric and photometric transforms. Inputs are resized to 640×640, and strong augmentations are disabled in the final epochs.

B. Performance Comparison

Table I reports the quantitative results on the DSLR, Smartphone, and DENTEX2023 datasets, together with parameter counts and FLOPs to highlight the accuracy–efficiency trade-off. Compared with the DEIM baseline, our model improves AP by **+1.3**, **+3.7**, and **+1.5** points on the DSLR, Smartphone, and DENTEX2023 datasets, respectively, while adding only $\sim$1.3M parameters and a small increase in FLOPs. The gains are especially pronounced on the Smartphone set, where severe blur, specular highlights, and framing artifacts make the task more challenging.

Qualitatively, as illustrated in Fig. 2, most baselines either miss ambiguous teeth or produce spurious boxes and duplicated FDI indices, especially for posterior teeth in blurred or over-exposed regions. Among them, only Dome-DETR is roughly comparable, whereas Faster R-CNN, CAF-YOLO-s, YOLOv11-s, RT-DETR-l, and DEIM show more missed detections and false positives, particularly on Smartphone and DENTEX2023. Our method yields tighter boxes and a coherent numbering sequence along each arch, with fewer missing posterior teeth and cross-arch confusions.

C. Ablation Studies

We perform ablation studies on the DENTEX2023 to isolate the contributions of HFSA, UniGate, and QGP under the same training setup as the main model.

Effect of HFSA, UniGate, and QGP. Table II summarizes the ablation results. Starting from the baseline, HFSA, UniGate, and QGP each yield consistent improvements, and combining all three modules achieves the best performance of 54.1 AP, 94.0 AP_{50}, and 55.9 AP_{75}, indicating that the HFSA-enhanced encoder and the UniGate–QGP decoder priors are complementary.

Effectiveness of HFSA. Adding HFSA improves AP from 52.6 to 53.2 and AP_{50} from 92.3 to 93.2. Fig. 3(b) shows that HFSA produces stronger and more continuous responses along the dental arch compared with the scattered activations of the baseline encoder. Fig. 4 further reveals that HFSA preserves more mid- to high-frequency energy ($0.1 < f < 0.4$) on the deepest FPN feature map, corresponding to tooth boundaries

TABLE I

PERFORMANCE COMPARISON ON THREE DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST RESULTS ARE UNDERLINED. PARAMS (M) AND FLOPS (G) ARE MEASURED AT 640×640.

Method	Year	Venue	Params (M)	FLOPs (G)	DSLR			Smartphone			DENTEX2023		
					AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Faster R-CNN [2]	2015	NeurIPS	41.61	178.0	67.8	90.9	79.9	33.5	66.9	29.4	38.7	85.6	26.4
CAF-YOLO-s [24]	2025	ICASSP	11.5	25.9	77.9	**97.8**	**91.2**	41.8	71.1	44.4	48.8	91.8	45.3
YOLOv8-s [25]	2024	arXiv	11.16	28.8	77.4	97.4	90.3	39.5	67.8	41.3	51.1	92.9	51.0
YOLOv11-s [26]	2024	arXiv	9.5	21.7	77.3	97.4	90.1	46.7	75.3	51.4	49.2	92.2	47.0
RT-DETR-l [27]	2024	CVPR	32.9	108.1	77.5	97.1	90.0	51.4	77.2	58.5	52.3	90.5	53.3
Dome-DETR [28]	2025	ACMMM	23.9	204.7	74.6	94.8	86.3	51.9	80.7	58.6	52.2	92.5	52.5
DEIM [17]	2025	CVPR	10.2	24.96	77.0	97.1	90.0	50.7	80.0	57.1	52.6	92.3	53.7
Ours (All)	2025	Ours	11.5	27.9	**78.3**	97.7	90.6	**54.4**	**83.9**	**60.7**	**54.1**	**94.0**	**55.9**

TABLE II

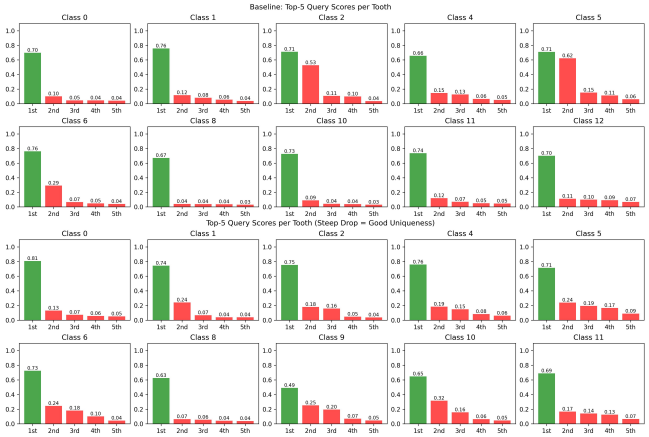
ABLATION OF HFSA, UNIGATE, AND QGP ON THE DENTEX2023 VALIDATION SET. ✓ DENOTES THAT THE COMPONENT IS ENABLED.

HFSA	UniGate	QGP	AP	AP_{50}	AP_{75}
			52.6	92.3	53.7
✓			53.2	93.2	54.2
	✓		53.1	93.7	54.9
		✓	53.0	93.9	52.7
✓	✓		53.5	93.9	53.2
✓		✓	53.1	93.5	53.4
	✓	✓	53.5	94.1	53.9
✓	✓	✓	**54.1**	**94.0**	**55.9**

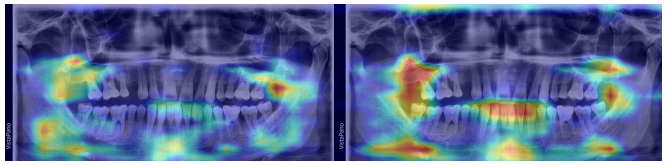
TABLE III

EFFECT OF ENABLING QGP ON DIFFERENT DECODER LAYERS ON DENTEX2023. L_1, L_2, AND L_3 DENOTE THE FIRST THREE DECODER LAYERS.

L_1	L_2	L_3	AP	AP_{50}	AP_{75}
✓			53.3	94.4	53.8
	✓		53.4	94.2	53.2
		✓	53.9	94.2	54.4
✓	✓	✓	53.4	94.0	54.4
✓	✓		53.6	94.2	53.3
	✓	✓	53.0	94.3	52.6
✓		✓	**54.1**	**94.0**	**55.9**



(a) Top-5 query scores with/without UniGate.



(b) Attention maps before/after HFSA.

Fig. 3. Qualitative effect of UniGate and HFSA on the DENTEX2023. (a) UniGate sharpens the query distribution so that each present tooth index is dominated by a single high-confidence query. (b) HFSA redistributes attention from gingiva and background to the dental arches and individual teeth, enhancing structural focus and suppressing specular highlights.

and fine details, which explains the improved localization of closely spaced crowns.

Effectiveness of QGP. Introducing QGP alone increases AP from 52.6 to 53.0 and AP_{50} from 92.3 to 93.9, and also contributes to the best overall performance when combined with HFSA and UniGate. As illustrated in Fig. 5, QGP makes query attention spatially coherent along the dental arch and constrains each query to focus on a short contiguous segment around its target tooth, rather than spreading across distant regions or both arches. Table III shows that enabling QGP on the first and third decoder layers (L_1 and L_3) gives the best trade-off, and we adopt this configuration by default.

Effectiveness of UniGate. Adding UniGate increases AP from 52.6 to 53.1 and AP_{75} from 53.7 to 54.9. Fig. 3(a) indicates that UniGate sharpens the query–tooth score distribution so that each tooth index is dominated by a single high-confidence query, reducing duplicate predictions and leading to more reliable detections.

IV. CONCLUSION

We presented a DETR-based framework for tooth detection and FDI numbering equipped with dentistry-aware priors. The Hybrid Frequency–Spatial Attention (HFSA) enhances robustness to imaging artifacts, whilst UniGate and the Query Graph Prior (QGP) enforce anatomical and indexing constraints. Experiments on three diverse datasets show consistent gains over baselines in both AP and FDI accuracy without increasing

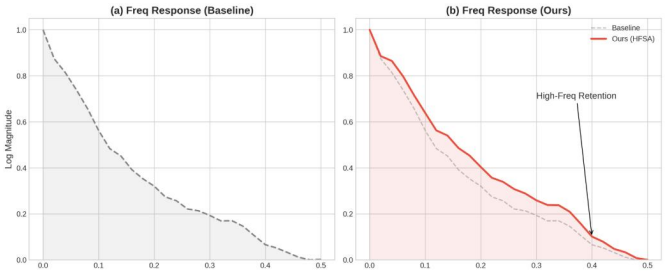


Fig. 4. Frequency analysis of encoder features on the DENTEX2023. The DEIM baseline (gray dashed) behaves like a strong low-pass filter on the deepest FPN feature map, while our HFSA encoder (red) preserves more mid-to-high frequency energy corresponding to tooth edges and fine structural details.

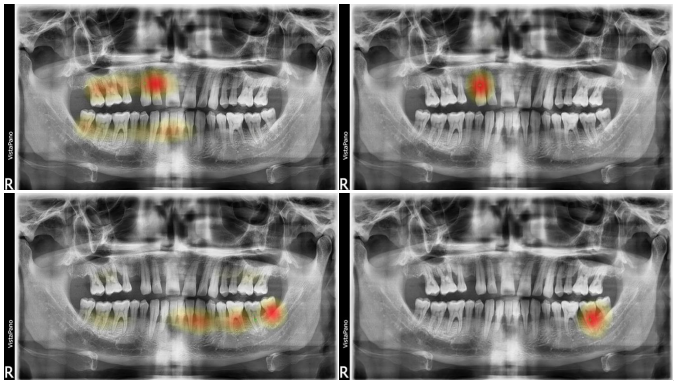


Fig. 5. Qualitative effect of QGP on query attention on the DENTEX2023 validation set. Left: the baseline model exhibits scattered attention that sometimes jumps across arches. Right: with QGP, attention is constrained to follow the dental arch and remains localized around neighboring teeth, reducing cross-arch confusion and local numbering swaps.

inference latency. This efficiency enables mobile deployment for teledentistry. Future work will extend this to video, 3D data, and pathology modeling.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 62571355).

REFERENCES

[1] ISO 3950:2016, "Dentistry – Designation system for teeth and areas of the oral cavity," International Organization for Standardization, 2016.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015, pp. 91–99.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot Multibox Detector," in *Proc. European Conf. Computer Vision (ECCV)*, 2016, pp. 21–37.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[5] H. Chen, H. Li, Y. Zhao, J. Zhao, and Y. Wang, "A Deep Learning Approach to Automatic Teeth Detection and Numbering Based on Object Detection in Dental Periapical Films," *Scientific Reports*, vol. 11, no. 1, p. 12034, 2021.

[6] S. Vinayahalingam, T. Xi, S. Bergé, T. Maal, and G. de Jong, "Automated Detection of Third Molars and Mandibular Nerve by Deep Learning," *Scientific Reports*, vol. 9, no. 1, p. 9007, 2019.

[7] A. F. Leite, K. de F. Vasconcelos, H. Willems, and R. Jacobs, "Artificial Intelligence-Driven Novel Tool for Tooth Detection and Segmentation on Panoramic Radiographs," *Clinical Oral Investigations*, vol. 25, no. 4, pp. 2257–2267, 2021.

[8] DENTEX 2023 Grand Challenge, "Tooth Recognition in Panoramic Radiographs," 2023. [Online]. Available: https://dentex.grand-challenge.org

[9] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Proc. European Conf. Computer Vision (ECCV)*, 2020, pp. 213–229.

[10] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[11] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic Anchor Boxes Are Better Queries for DETR," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2022.

[12] Y. Wang, X. Zhang, T. Yang, and J. Sun, "Anchor DETR: Query Design for Transformer-Based Detector," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2567–2575.

[13] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR Training by Introducing Query Denoising," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13619–13627.

[14] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with Improved Denoising Anchor Boxes for End-to-End Object Detection," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023.

[15] Q. Chen, X. Chen, J. Wang, S. Zhang, K. Yao, H. Feng, J. Han, E. Ding, G. Zeng, and J. Wang, "Group DETR: Fast DETR Training with Group-wise One-to-Many Assignment," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2023, pp. 6633–6642.

[16] D. Jia, Y. Yuan, H. He, X. Wu, H. Yu, W. Lin, L. Sun, C. Zhang, and H. Hu, "DETRs with Hybrid Matching," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19702–19712.

[17] S. Huang, Z. Lu, X. Cun, Y. Yu, X. Zhou, and X. Shen, "DEIM: DETR with Improved Matching for Fast Convergence," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2025.

[18] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu, "D-FINE: Redefine Regression Task in DETRs as Fine-Grained Distribution Refinement," in *Int. Conf. Learning Representations (ICLR)*, 2025.

[19] A. Wirtz, S. G. Mirashi, and S. Wesarg, "Automatic Teeth Segmentation in Panoramic X-Ray Images Using a Coupled Shape Model in Combination with a Neural Network," in *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018, pp. 712–719.

[20] S. Tian, M. Wang, N. Dai, H. Ma, L. Li, L. Fiorenza, Y. Sun, Y. Lee, D. Guatelli-Steinberg, and H. Hua, "Automatic Classification and Segmentation of Teeth on 3D Dental Model Using Hierarchical Deep Learning Networks," *IEEE Access*, vol. 7, pp. 84817–84828, 2019.

[21] J.-H. Lee, D.-H. Kim, S.-N. Jeong, and S.-H. Choi, "Graph Convolutional Networks for Teeth Identification in Dental Panoramic Radiographs," *Scientific Reports*, vol. 12, no. 1, p. 14223, 2022.

[22] G. Mena, D. Belanger, S. Linderman, and J. Snoek, "Learning Latent Permutations with Gumbel-Sinkhorn Networks," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

[23] K. S. Siddartha, "Dental Anatomy Dataset - Yolov8," *Kaggle*, 2024. [Online]. Available: https://www.kaggle.com/datasets/saisiddartha69/dental-anatomy-dataset-yolov8

[24] Z. Chen and S. Lu, "CAF-YOLO: A Robust Framework for Multi-Scale Lesion Detection in Biomedical Imagery," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[25] G. Jocher and A. Chaurasia, "Ultralytics YOLOv8," software, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[26] G. Jocher and J. Qiu, "Ultralytics YOLO11," software, version 11.0.0, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[27] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs Beat YOLOs on Real-time Object Detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16955–16964.

[28] Z. Hu, P. Wu, J. Chen, H. Zhu, Y. Wang, Y. Peng, H. Li, and X. Sun, "Dome-DETR: DETR with Density-Oriented Feature-Query Manipulation for Efficient Tiny Object Detection," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 101–110.