GQD-DETR: A Geometry-Aware and Quality-Adaptive Transformer for Robust Tooth Detection in Heterogeneous Dental Images

Chen Zhang*, 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*—Tooth position detection underpins automated analysis of oral images, yet existing methods often fail on low-quality radiographs with blur, overlapping teeth, and low contrast. We propose GQD-DETR, an encoder–decoder framework tailored for robust tooth localization under real-world imaging conditions. GQD-DETR comprises three key components: (1) a Geometry-Aware Attentive High-Resolution module that combines polar coordinate priors with a high-resolution attention pathway to capture dental-arch curvature and fine details; (2) a Quality-Adaptive Cross-Scale Interaction module that performs quality-aware feature filtering and cross-scale context aggregation to handle noise and illumination inhomogeneity; and (3) a Discriminative Anisotropic Decoding module with learnable anisotropic positional encoding and query channel attention, which stabilizes query refinement in crowded regions and improves discrimination between spatially adjacent, visually similar teeth. Experiments on our private dataset and the public DENTEX2023 benchmark show that GQD-DETR achieves 58.98 mAP and 54.67 mAP, outperforming strong baselines by 1.62 and 1.43 mAP, respectively, and effectively handling multi-source images including panoramic radiographs, smartphone photos, and DSLR images.**

***Index Terms*—Dental image analysis, tooth detection, object detection, transformer, deep learning, medical image processing, panoramic radiography**

I. Introduction

Automatic tooth detection and classification in dental images is a fundamental task in digital dentistry, supporting applications from orthodontic planning to computer-aided diagnosis [1], [2], [3], [4]. With the increasing adoption of digital imaging, clinicians now work with heterogeneous data, ranging from standard panoramic radiographs to intraoral cameras and smartphone photographs [5], [4], [6], [7]. In real-world practice, these images often exhibit motion blur, low contrast, uneven illumination, and metal artifacts [1], [8], [9], [10]. Combined with the small size, dense arrangement, and irregular spatial distribution of teeth, these degradations make robust automatic detection particularly challenging and limit the reliability of current tools.

Most existing deep-learning approaches for tooth detection follow the CNN-based object detection paradigm. Early works adapted Faster R-CNN [11] and YOLO variants [12] to enumerate teeth on curated radiograph datasets [2], [3], [5]. While effective on high-quality images, these convolution-based detectors degrade on blurred or artifact-corrupted inputs, where they have difficulty separating adjacent teeth and suppressing background structures, resulting in false positives and inaccurate localization. Recent works exploring segmentation pipelines [13] and improved CNN architectures show progress, but their robustness on heterogeneous, lower-quality clinical data remains limited.

Transformer-based object detectors such as DETR [14] and its variants (e.g., Deformable DETR [15], DINO [16], RT-DETR [17]) formulate detection as set prediction and use self-attention to capture global context, which is attractive for resolving local ambiguities in degraded images. However, standard DETR-style architectures are designed for general scenarios and lack explicit mechanisms to deal with the noise and uncertainty of low-quality medical images. They typically process all features uniformly, without dynamically filtering artifacts or enhancing weak signals in blurred regions, and rely on isotropic positional priors that do not exploit the structural regularity of the dental arch. As a result, when applied to dental images with poor quality and crowded targets, these Transformer detectors often produce unstable tooth queries and inconsistent predictions.

To address these challenges, we propose **GQD-DETR**, a unified encoder–decoder framework specifically tailored for robust tooth localization under real-world imaging conditions. GQD-DETR integrates geometry-aware priors and quality-adaptive mechanisms into a DETR-style architecture. In the encoder, a Geometry-Aware Attentive High-Resolution (GAHR) module exploits polar coordinate priors and a high-resolution attention pathway to recover structural cues when visual boundaries are indistinct. A Quality-Adaptive Cross-Scale Interaction (QACI) module dynamically evaluates feature quality to suppress noise from artifacts and aggregates multi-scale context to handle illumination inhomogeneity. In the decoder, a Discriminative Anisotropic Decoding (DAD) module with learnable anisotropic positional encoding and query channel attention enhances the discrimination of spatially adjacent, visually similar teeth,

stabilizing localization in low-contrast, crowded regions.

The contributions of this work are summarized as follows:

- We propose GQD-DETR, a robust tooth detection framework designed for low-quality and heterogeneous dental images (e.g., blur, low contrast, multi-source acquisition) and for the inherent challenges of small, crowded, irregularly distributed teeth.
- We introduce the Geometry-Aware Attentive High-Resolution (GAHR) module and the Quality-Adaptive Cross-Scale Interaction (QACI) module in the encoder to preserve dental-arch structure via polar priors and to perform quality-aware noise suppression and cross-scale context aggregation.
- We design a Discriminative Anisotropic Decoding (DAD) module equipped with learnable anisotropic positional encoding and query channel attention, which yields more discriminative tooth queries and stable localization, enabling GQD-DETR to surpass strong DETR-style baselines on varying-quality datasets, including panoramic radiographs and smartphone photos.

II. Method

A. Overall Architecture

As shown in Fig. 1, we propose GQD-DETR, a DETR-style encoder–decoder tailored for robust tooth detection addressing two main multi-source data challenges: poor image quality and complex curvilinear dental arches. First, a CNN extracts multi-scale features $\{\mathbf{C}_s\}_{s\in\{8,16,32\}}$. These are refined by a hybrid encoder using the GAHR module to preserve fine-grained geometry and the QACI module to explicitly mitigate image degradation. Subsequently, an NMS-free transformer decoder, enhanced by the DAD module to exploit the arch's anisotropic nature, iteratively refines object queries. The final detection output is:

$$\mathcal{Y} = \left\{(\hat{\mathbf{b}}_i, \hat{c}_i)\right\}_{i=1}^{N_q},$$

where N_q queries yield predicted bounding boxes $\hat{\mathbf{b}}_i$ and tooth categories $\hat{c}_i$.

B. Hybrid Encoder with GAHR and QACI

The hybrid encoder is designed to process multi-scale backbone features $\{\mathbf{C}_s\}_{s\in\{8,16,32\}}$ through two specialized modules that handle the unique geometry of the dental arch and the substantial variance in X-ray image quality. Before entering these modules, features from different scales are projected to a uniform hidden dimension D via a projection layer f_{proj}, facilitating consistent interaction across levels.

1) Geometry-Aware Attentive High-Resolution (GAHR): The highest-resolution feature map ($\mathbf{C}_8$) is critical for detecting small dental structures but is often susceptible to spatial noise. GAHR enhances these features through a dual-pathway mechanism that combines explicit geometric priors with semantic attention.

Recognizing that teeth are arranged along a curved arch rather than a rectilinear grid, the *geometry path* integrates polar-enhanced spatial priors. By explicitly encoding the radial distance $\mathbf{R}_{ij}$ and angular position $\mathbf{\Theta}_{ij}$ relative to the image center, the model gains rotation-invariant geometric cues alongside standard Cartesian coordinates $\mathbf{X}$, $\mathbf{Y}$ [18]:

$$\mathbf{E}_{\text{geom}} = f_{\text{polar}}\left(\text{Concat}\left[\mathbf{C}_8; \mathbf{X}; \mathbf{Y}; \mathbf{R}; \mathbf{\Theta}\right]\right), \tag{1}$$

where $\mathbf{R}_{ij} = \sqrt{x_{ij}^2 + y_{ij}^2}$ is the radial distance and $\mathbf{\Theta}_{ij} = \arctan 2(y_{ij}, x_{ij})$ is the angular position relative to the image center.

Simultaneously, the *high-resolution path* focuses on feature refinement. It applies sequential channel attention ($\mathcal{A}_c$) to highlight informative feature maps and spatial attention ($\mathcal{A}s$) [19] to focus on tooth boundaries, effectively suppressing background noise:

$$\mathbf{E}\text{hr} = \mathcal{A}s\left(\mathcal{A}c\left(f\text{conv}(\mathbf{C}_8)\right)\right), \tag{2}$$

where $\mathcal{A}c$ is the channel attention and $\mathcal{A}s$ is the spatial attention mechanism. The outputs $\mathbf{E}$geom and $\mathbf{E}$hr are then fused, combining geometric structural guidance with attentive semantic features to form the final enhanced representation.

2) Quality-Adaptive Cross-Scale Interaction (QACI): To mitigate common radiographic degradations (e.g., motion blur, artifacts), QACI dynamically processes semantically rich, lower-resolution features ($\mathbf{C}_{16}, \mathbf{C}_{32}$).

The *Quality-Aware Module (QAM)* addresses this by aggregating three complementary branches: denoising (f_{dn}), balancing (f_{bal}), and detail-preserving (f_{dt}). A lightweight estimator analyzes global statistics to predict adaptive weights, dynamically fusing these features:

$$\mathbf{E}^{\text{Q}} = \sum_{k\in\{\text{dn, bal, dt}\}} w_k \cdot f_k(\mathbf{E}), \tag{3}$$

where $w_k \in \mathbb{R}^3$ denotes the predicted adaptive weights, and $f_k(\mathbf{E})$ represents the feature from branch k.

To expand individual receptive fields, features from each scale are average- and max-pooled into global tokens to capture representative statistics. Multi-Head Self-Attention (MHSA) then efficiently exchanges long-range context across scales:

$$[\mathbf{T}'_8, \mathbf{T}'_{16}, \mathbf{T}'_{32}] = \text{MHSA}\left(\text{Concat}[\mathbf{T}_8, \mathbf{T}_{16}, \mathbf{T}_{32}]\right), \tag{4}$$

where $\mathbf{T}_s$ and $\mathbf{T}'_s$ are the pooled and MHSA-refined global tokens, respectively. Finally, $\mathbf{T}'_s$ are fused back into local feature maps via a learnable gate.

C. Discriminative Anisotropic Decoding (DAD)

Tooth position detection exhibits an anisotropic spatial structure, with teeth arranged mainly along a horizontal arch. Standard DETR decoders rely on isotropic positional encodings, which limits their ability to model such geometry. The proposed DAD module introduces *Learnable Anisotropic Positional Encoding (Anisotropic PE)* and *Query Channel Attention (QCA)* to improve geometric alignment and query discrimination.

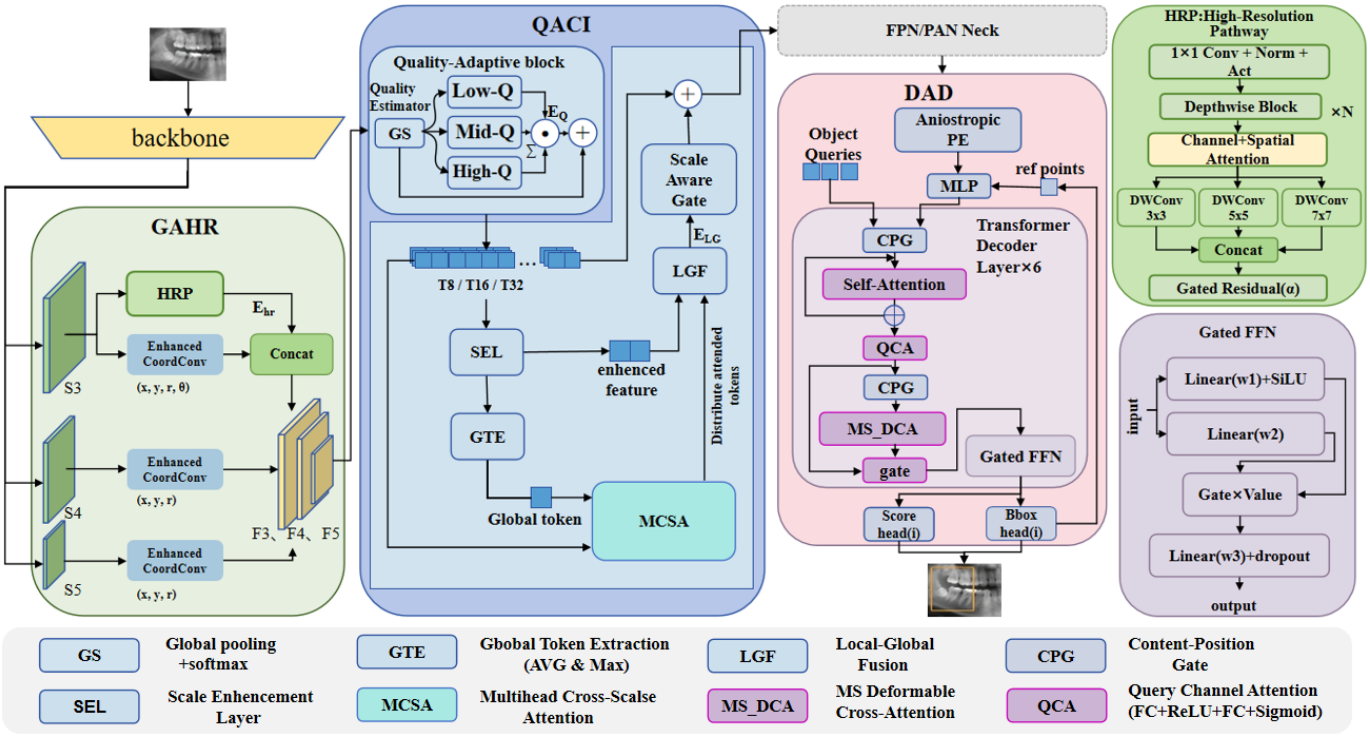

Fig. 1. Overview of the GQD-DETR architecture. The backbone extracts multi-scale features, which are processed by the hybrid encoder with the GAHR and QACI modules. The decoder, enhanced by the DAD module, refines queries for robust tooth localization under varying image conditions.

1) Learnable Anisotropic Positional Encoding (Anisotropic PE): Each decoder query is associated with a normalized reference point (c_i^x, c_i^y). To adapt to the horizontal dominance of dental arches, we introduce learnable axis-specific scales λ_x and λ_y:

$$\mathrm{PE}_i^{\text{aniso}} = \left[\Phi(c_i^x \cdot \lambda_x, D/2);\ \Phi(c_i^y \cdot \lambda_y, D/2)\right], \tag{5}$$

where $\Phi(u, f)$ is the sine–cosine encoding function and D is the embedding dimension. The positional embedding is then projected as:

$$\mathbf{q}_i^{\text{pos}} = \mathrm{MLP}_{\text{pos}}(\mathrm{PE}_i^{\text{aniso}}), \tag{6}$$

where λ_x and λ_y are learnable scalars that modulate horizontal and vertical frequency bandwidths.

2) Query Channel Attention (QCA) and Gated Refinement: To enhance discriminability among spatially adjacent teeth, QCA reweights channel responses after the self-attention update:

$$\mathbf{q}_i^{\text{att}} = \mathbf{q}_i^{\text{content}} \odot \sigma\left(\mathbf{W}_2\, \delta(\mathbf{W}_1 \mathbf{q}_i^{\text{content}})\right), \tag{7}$$

where $\odot$ is element-wise multiplication, σ is the sigmoid activation, δ is ReLU, and $\mathbf{W}_i$ are weights of a bottleneck MLP.

We further adopt a *Gated FFN* to filter information flow:

$$\mathbf{q}_i^{\text{out}} = \mathbf{W}_{\text{out}}\left(\mathrm{SiLU}(\mathbf{W}_{\text{g}} \mathbf{q}_i^{\text{att}}) \odot (\mathbf{W}_{\text{v}} \mathbf{q}_i^{\text{att}})\right), \tag{8}$$

where $\mathbf{W}_{\text{g}}, \mathbf{W}_{\text{v}}, \mathbf{W}_{\text{out}}$ are linear projections. This gating works jointly with QCA to promote informative channels and suppress noise, improving detection in crowded regions.

D. Training and Inference

GQD-DETR is trained end-to-end using bipartite matching. The objective function combines Focal Loss for classification with L1 and GIoU losses for box regression. During inference, object queries are directly decoded into detections without Non-Maximum Suppression (NMS). The proposed modules incur minimal computational overhead, maintaining efficient real-time performance.

III. Experiments

A. Datasets and Splits

We evaluate on three 2D dental image datasets to assess cross-domain robustness: **(1) Private-Phone**: smartphone images with heterogeneous illumination, motion blur, and occlusions, annotated with tooth bounding boxes and dental indices; **(2) DLSR** [25]: a curated public photographic dataset from dedicated intra/extra-oral cameras; and **(3) DENTEX 2023**: a panoramic radiograph benchmark [8]. Unless otherwise specified, we use each dataset's official or commonly adopted train/validation/test split,[1] train all methods on the training split, and evaluate on the held-out test split without test-time augmentation.

B. Evaluation Protocol

We report COCO-style detection metrics: $\mathrm{mAP}_{50:95}$, AP_{50}, and class-agnostic Precision/Recall. For efficiency, we also

[1]For *Private-Phone*, we adopt a fixed patient-level split to avoid identity leakage.

TABLE I
COMPARISON WITH DEIM AND RECENT DETECTORS ON THREE DENTAL IMAGE SOURCES. AP DENOTES MAP@0.5:0.95. WE ALSO REPORT AP_{50} AND AP_{75} FOR EACH DATASET.

Method	Params (M)↓	FLOPs (G)↓	Private-Phone			DLSR			DENTEX 2023		
			AP↑	AP_{50} ↑	AP_{75} ↑	AP↑	AP_{50} ↑	AP_{75} ↑	AP↑	AP_{50} ↑	AP_{75} ↑
Faster R-CNN [11]	41.6	178	35.48	69.88	34.50	67.66	91.02	79.93	38.66	85.54	26.51
YOLOv8m [20]	26.2	80.6	54.08	84.27	59.81	77.77	97.34	90.00	48.78	91.07	45.94
YOLOv11m [21]	20.1	68.0	54.98	84.05	62.18	77.91	97.33	90.61	48.75	91.09	44.94
CAF-YOLO [22]	26.6	73.6	51.26	82.40	58.08	77.68	97.82	89.91	48.58	91.46	44.37
RT-DETR-R101 [17]	76	259	49.21	81.53	54.88	77.44	97.42	89.93	47.74	84.78	47.78
Dome-DETR [23]	23.9	284.5	55.46	84.12	61.48	77.75	97.90	90.71	51.36	91.11	50.37
DEIM [24]	**19.2**	**56.4**	57.36	89.07	64.42	77.82	98.03	**91.86**	53.24	94.28	53.67
Ours	21.6	66.1	**58.98**	**91.24**	**66.83**	**78.53**	**98.05**	91.01	**54.67**	**94.80**	**56.22**

TABLE II
ABLATIONS ON THE *Dentex 2023* SET. AP = $MAP_{50:95}$.

GAHR	QACI	DAD	AP↑	AP_{50} ↑	AP_{75} ↑	Params (M)↓
			53.24	94.28	53.67	—
✓			53.89	94.21	54.63	+∼0.3
	✓		53.76	94.26	54.64	+∼1.8
		✓	53.77	93.69	53.95	+∼0.3
✓	✓		54.17	94.08	54.00	+∼2.1
✓		✓	53.95	94.35	53.75	+∼0.6
	✓	✓	53.92	94.69	54.17	+∼2.1
✓	✓	✓	**54.67**	**94.80**	**56.22**	+∼2.4

TABLE III
INTERNAL ABLATION OF THE GAHR MODULE ON *DENTEX 2023*. THE FIRST ROW REPRESENTS THE MODEL PERFORMANCE WHEN THE GAHR MODULE IS REMOVED (OR REPLACED BY IDENTITY).

Variant	AP	AP_{50}	Params (M)
None (w/o GAHR)	53.92	94.69	—
CoordConv Only	54.18	94.07	+∼0.2
High-Res Pathway Only	54.15	94.18	+∼0.1
Full GAHR	**54.67**	**94.80**	+∼0.3

report FLOPs and parameter counts (M) for each model. All results are single-scale, single-model unless noted.

C. Implementation Details

We implement our model based on the DEIM framework and integrate the three proposed modules (GAHR, QACI, and DAD). Unless otherwise stated, we follow the optimization and training schedule from DEIM [24]. The transformer backbone uses a hidden dimension of $d = 256$, 8 attention heads, and 6 decoder layers with distribution-based box refinement (reg_max$= 32$). The hybrid encoder processes three feature pyramid levels with strides $\{8, 16, 32\}$. We employ BatchNorm during training and apply BN–Conv fusion in deploy mode for re-parameterizable convolutional blocks. Input images are resized to 640×640 pixels. For data augmentation, we use standard color/contrast adjustments and mild geometric transformations; no synthetic copy–paste augmentation is applied. All experiments are conducted on two NVIDIA RTX 5090 GPUs with fixed random seeds for reproducibility.

D. Main Results

Table I details the quantitative results, where our GQD-DETR consistently achieves superior performance. Specifically, it attains a $\mathrm{mAP}_{50:95}$ of 58.98 on *Private-Phone* (surpassing DEIM by 1.62), alongside 78.53 on *DLSR* and 54.67 on *DENTEX 2023*, yielding respective improvements of 0.71 and 1.43. These results demonstrate robust, competitive performance against state-of-the-art methods across heterogeneous conditions. Qualitatively (Fig. 2), GQD-DETR exhibits consistently better predictions than competing methods. It effectively avoids common false detections and successfully identifies teeth absent from ground-truth annotations.

E. Ablation Study

a) Impact of proposed modules: Table II presents a component-wise ablation study on the DENTEX 2023 dataset.Relative to the baseline setting of 53.24 mAP, the incorporation of GAHR alone elevates performance to 53.89 mAP. This gain of **0.65** highlights the benefit of high-resolution features in preserving dental details.Subsequently, the addition of QACI and DAD contributes further increases of 0.52 and 0.53 AP, respectively. Their pairwise combinations yield approximately additive improvements.The integration of all three modules culminates in the best result of **54.67** mAP, suggesting that geometry preservation, quality-adaptive fusion, and discriminative decoding function as complementary mechanisms.

b) Internal analysis of GAHR: We perform an internal analysis on DENTEX 2023, detailed in Table III, to isolate the contributions of the GAHR branches. Individual activation of the geometry-aware branch and the high-resolution pathway improves the baseline mAP of 53.92 to 54.18 and 54.15, respectively. However, the full configuration delivers the highest performance of **54.67 mAP**. This overall improvement of 0.75 mAP demonstrates that the two branches offer complementary inductive biases, surpassing the capabilities of either branch operating in isolation. Grad-CAM visualizations in Fig. 3 further validate this design. While the baseline and single-branch models often exhibit scattered or locally restricted activations, the *Full GAHR* setting generates compact and continuous heatmaps. As shown in Fig. 3a, this combined

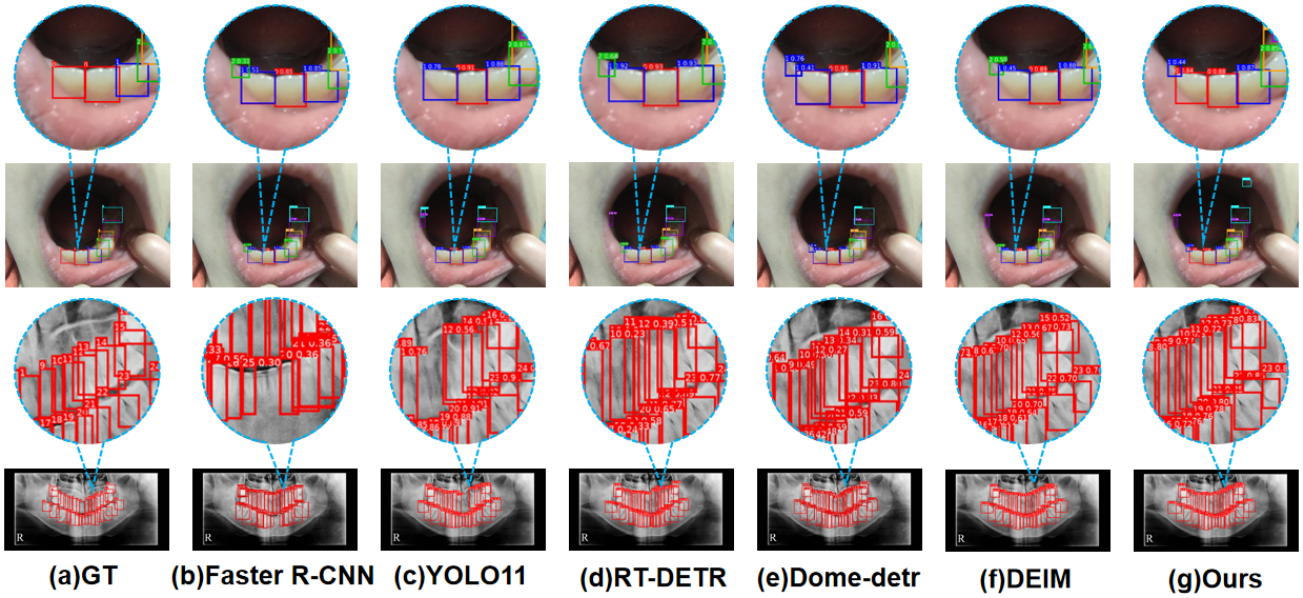

Fig. 2. Comparison of inference results across different models on the dataset.

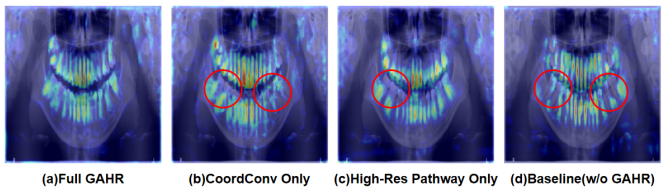

Fig. 3. Grad-CAM visualizations of GAHR ablation settings.Panel (a) displays the Full GAHR; panel (b) shows CoordConv Only; panel (c) represents the High-Res Pathway Only; and panel (d) illustrates the Baseline without GAHR.The full module in panel (a) exhibits the most focused activations on tooth regions compared to the single-branch variants in (b) and (c) or the baseline in (d).

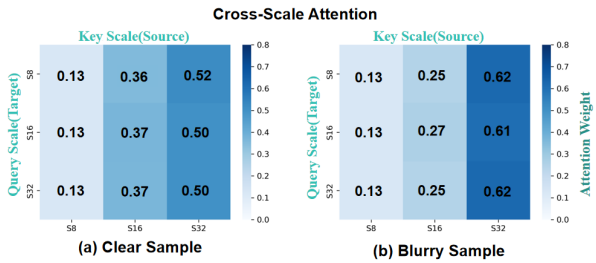

Fig. 4. Cross-scale attention maps demonstrating QACI's instance-aware adaptability. Unlike the balanced feature distribution in (a), Sample (b) significantly shifts focus towards global context (S32) to resolve input ambiguity.

approach precisely targets tooth structures while suppressing background noise, confirming the effective synergy between geometric priors and high-resolution features.

c) Quality-adaptive processing effectiveness: To further validate the QACI module on the Private-Phone dataset, which features varying image qualities, we visualize its internal cross-scale attention maps in Figure 4. For the **Clear Sample (a)**, the model maintains a balanced attention between mid-level and high-level features. Conversely, for the **Blurry Sample (b)**, the focus significantly shifts towards the global context to resolve local ambiguity, with the S32 weight rising to ≈ 0.62. This visualization confirms that QACI performs instance-aware dynamic recalibration rather than static fusion. This adaptive capability effectively explains the substantial performance gain reported in Table II: the module's ability to leverage global context for low-quality inputs boosts the model's robustness, improving the performance.

d) Discriminative decoding analysis: The DAD module significantly enhances localization (AP_{75}). Compared to the baseline's scattered t-SNE projections, DAD yields more compact query position embedding clusters with a smoother gradient for horizontal locations c_x (Fig. 5(a,b)), explicitly encoding tooth-arch geometry. Furthermore, DAD transforms the baseline's fragmented self-attention maps into a wider, continuous near-diagonal band (Fig. 5(c,d)), reflecting stronger interactions and restored local connectivity among neighboring teeth. These enhanced spatial coherence and discriminative representations—driven by DAD's learnable anisotropic positional encoding and query channel attention—explain the AP_{75} gains and robustness in crowded dental arrangements.

IV. Conclusion

We presented GQD-DETR, a robust Transformer-based framework tailored for tooth position detection across heterogeneous 2D dental images. To address the challenges of low-quality imaging and complex dental anatomy, we introduced three targeted enhancements. In the encoder, the Geometry-Aware Attentive High-Resolution (GAHR) module integrates polar priors to preserve fine structural details, while the Quality-Adaptive Cross-Scale Interaction (QACI) module dynamically filters noise and aggregates global context to

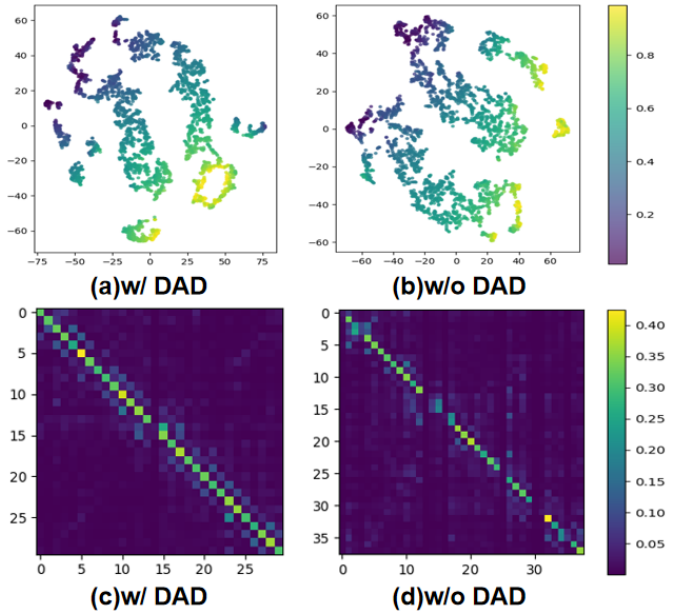

Fig. 5. Visualization of the effect of the proposed DAD module.(a)-(b) t-SNE (only tooth queries, sorted by x-coordinate) visualizations of feature representations without and with DAD, respectively.(c)-(d) Self-attention heatmaps without and with DAD, respectively (axes denote spatially sorted tooth query indices).

mitigate image degradation. In the decoder, the Discriminative Anisotropic Decoding (DAD) module leverages learnable positional encodings and query channel attention to stabilize localization and distinguish similar teeth. Extensive experiments demonstrate that GQD-DETR consistently outperforms state-of-the-art baselines on diverse datasets, confirming that explicitly modeling geometric priors and quality adaptability is key to reliable automated dental analysis in real-world clinical scenarios.

ACKNOWLEDGMENT

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

REFERENCES

[1] S. Sadr, R. Rokhshad, Y. Daghighi, M. Golkar, F. Tolooie Kheybari, F. Gorjinejad, A. Mataji Kojori, P. Rahimirad, P. Shobeiri, M. Mahdian, and H. Mohammad-Rahimi, "Deep learning for tooth identification and numbering on dental radiography: a systematic review and meta-analysis," *Dentomaxillofacial Radiology*, vol. 53, no. 1, pp. 5–21, 2024.

[2] D. V. Tuzoff, L. N. Tuzova, M. M. Bornstein, A. S. Krasnov, M. A. Kharchenko, S. I. Nikolenko, M. M. Sveshnikov, and G. B. Bednenko, "Tooth detection and numbering in panoramic radiographs using convolutional neural networks," *Dentomaxillofacial Radiology*, vol. 48, no. 4, p. 20180051, 2019.

[3] M. Estai, M. Tennant, D. Gebauer, A. Brostek, J. Vignarajan, M. Mehdizadeh, and S. Saha, "Deep learning for automated detection and numbering of permanent teeth on panoramic images," *Dentomaxillofacial Radiology*, vol. 51, no. 2, p. 20210296, 2022.

[4] M. Bonfanti-Gris, A. Herrera, M. P. Salido Rodríguez-Manzaneque, F. Martínez-Rus, and G. Pradíes, "Deep learning for tooth detection and segmentation in panoramic radiographs: a systematic review and meta-analysis," *BMC Oral Health*, vol. 25, p. 1280, 2025.

[5] E. Bilgir, İ. Ş. Bayrakdar, Ö. Çelik, K. Orhan, F. Akkoca, H. Sağlam, A. Odabaş, A. F. Aslan, C. Ozcetin, M. Kıllı, and I. Rozylo-Kalinowska, "An artificial intelligence approach to automatic tooth detection and numbering in panoramic radiographs," *BMC Medical Imaging*, vol. 21, p. 124, 2021.

[6] M. T. G. Thanh, N. Van Toan, V. T. N. Ngoc, N. T. Tra, C. N. Giap, and D. M. Nguyen, "Deep learning application in dental caries detection using intraoral photos taken by smartphones," *Applied Sciences*, vol. 12, no. 11, p. 5504, 2022.

[7] M. Estai, Y. Kanagasingam, M. Tennant, and S. Bunt, "A systematic review of the research evidence for the benefits of teledentistry," *Journal of Telemedicine and Telecare*, vol. 24, no. 3, pp. 147–156, 2018.

[8] I. E. Hamamci, S. Er, E. Simsar, A. E. Yuksel, S. Gultekin, S. D. Ozdemir, K. Yang, H. B. Li, S. Pati, B. Stadlinger, *et al.*, "Dentex: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic x-rays," *arXiv preprint arXiv:2305.19112*, 2023.

[9] M. Xu, Y. Wu, Z. Xu, P. Ding, H. Bai, and X. Deng, "Robust automated teeth identification from dental radiographs using deep learning," *Journal of Dentistry*, vol. 136, p. 104607, 2023.

[10] C. M. Hyun, T. Bayaraa, H. S. Yun, T.-J. Jang, H. S. Park, and J. K. Seo, "Deep learning method for reducing metal artifacts in dental cone-beam ct using supplementary information from intra-oral scan," *Physics in Medicine & Biology*, vol. 67, no. 17, p. 175007, 2022.

[11] 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*, 2015.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[13] A. F. Leite, A. Van Gerven, H. Willems, T. Beznik, P. Lahoud, H. Gaêta-Araujo, *et al.*, "Artificial intelligence-driven novel tool for tooth detection and segmentation on panoramic radiographs," *Clinical Oral Investigations*, vol. 25, no. 4, pp. 2257–2267, 2021.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020.

[15] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021.

[16] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "Dino: DETR with improved denoising anchor boxes for end-to-end object detection," in *International Conference on Learning Representations*, 2023.

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[18] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

[20] G. Jocher and A. Chaurasia, "Ultralytics yolov8." https://github.com/ultralytics/ultralytics, 2023. Software.

[21] G. Jocher and J. Qiu, "Ultralytics yolo11." https://github.com/ultralytics/ultralytics, 2024. Software.

[22] Z. Chen and S. Lu, "Caf-yolo: A robust framework for multi-scale lesion detection in biomedical imagery," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2025.

[23] 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*, pp. 101–110, ACM, 2025.

[24] S. Huang, Z. Lu, X. Cun, Y. Yu, X. Zhou, and X. Shen, "Deim: DETR with improved matching for fast convergence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[25] K. S. Siddartha, "Dental anatomy dataset - yolov8." https://www.kaggle.com/datasets/saisiddartha69/dental-anatomy-dataset-yolov8, 2024.