MASKED ALIGNMENT PRETRAINING FROM LOCAL TO GLOBAL FOR INTRAORAL IMAGE REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised pretraining for intraoral photographs remains underexplored because generic objectives overlook the specific anatomy of the dental arch (*e.g.* bilaterally symmetrical and grouped into six sextants). We present a masked alignment pretraining framework for dental-arch-aware representation learning based on intraoral images. Random, single-tooth, and sextant crops are aligned with dense features at corresponding locations in full-mouth images using geometric overlap as supervision. Masked reconstruction of these crops further captures tooth texture and morphology that complement structural cues. FDI notation (the tooth numbering system) and tooth anatomical relations additionally regularize the tooth representation, so that mirror-symmetric or adjacent teeth, or those in the same sextant are closer, while distant teeth are separated. We also introduce YoungDent, a five-view intraoral dataset with over 5K images from more than 1,000 adolescents captured under non-clinical settings, and combine it with public data into a 10K-scale corpus. Compared with generic SSL baselines, our dental-arch-aware pretraining better captures tooth anatomical organization and improves representation quality, showing clear potential as a foundation for downstream tooth segmentation tasks.

1 INTRODUCTION

Self-supervised learning (SSL) enables transferable visual representation learning from large collections of unlabeled images (Siméoni et al., 2025; Chen et al., 2020; 2021; Chen & He, 2021; He et al., 2022a; Assran et al., 2023). Its success in natural-image understanding has motivated its adoption in medical imaging, where clinical image annotations are costly and require clinical expertise (Huang et al., 2023a; Moutakanni et al., 2026; Kim et al., 2026; Yao et al., 2025). In dental imaging, however, representation learning remains largely task-specific, particularly for tooth segmentation, and often relies on specialized architectures not designed for cross-task transfer (Cui et al., 2021; Xiong et al., 2023; Xi et al., 2025). Recent SSL methods for dental images mainly target cone-beam computed tomography (CBCT) or intraoral scans (IOS) (He et al., 2022b; Almalki & Latecki, 2024; Son et al., 2025), both of which require dedicated equipment and are difficult to scale beyond clinics.

Intraoral photographs offer a more accessible basis for foundation dental representation learning because they can be captured with commodity cameras or smartphones. Yet existing systems are typically developed separately for individual diagnostic or segmentation tasks (Zou et al., 2024; Liu et al., 2024), leaving a central question unanswered: can a single image encoder learn transferable dental representations from intraoral photographs? Currently, directly applying generic SSL methods such as SimCLRv2 (Chen et al., 2020), DINOv2 (Oquab et al., 2023), or MAE (He et al., 2022a) overlooks two properties of this domain. First is the difficulty in capturing fine-grained clinical differences. Teeth share similar color and texture but differ subtly in crown morphology, geometry, and anatomical context, while also needing to be distinguished from surrounding oral tissues. Second is the overlooked dental-arch structure. The dental arch has a consistent cross-subject structure. Tooth identity is tied to position in the dental arch, and contralateral, adjacent, and morphologically corresponding teeth follow predictable relations. Generic objectives based on view invariance or masked reconstruction do not explicitly capture this ordered anatomy. Nor do they reflect the coarse-to-fine way in which dentists examine a full dentition, a sextant, and individual teeth. Accordingly,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

local representations should correspond to the correct global regions, remain stable across views, and preserve anatomical relations among teeth. In summary, the core gap in SSL for dental images is the lack of structure-aware pretraining objectives capable of encoding the ordered anatomy of the dental arch, tooth relations, and fine-grained clinical semantics.

We therefore propose a masked alignment pretraining framework for dental-arch-aware representation learning from intraoral photographs. Inspired by position-based SSL methods such as VoCo (Wu et al., 2024), which predict the overlap between crops and full images in 3D medical imaging, we align random, single-tooth, and sextant crops with dense patch features at their corresponding locations in a full-mouth image. This geometric overlap provides the self-supervised target, grounding local features in the global dentition. Because spatially separate teeth may still be anatomically related, we use the FDI tooth-numbering system to reduce negative alignment penalties between related teeth and regularize dense features according to tooth identity, symmetry, adjacency, and morphology. Furthermore, a masked reconstruction branch retains complementary appearance cues, while masked-clean feature consistency improves robustness to partial observations. These objectives are integrated into a momentum teacher–student framework.

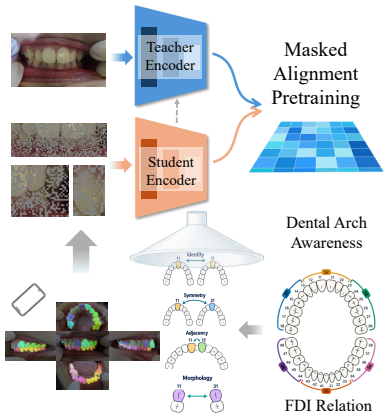


Figure 1: Overview of masked alignment pretraining. Masked crops are aligned with full-mouth context, while FDI priors organize tooth representations.

We further construct *YoungDent*, comprising more than 5,000 smartphone-captured intraoral photographs from over 1,000 adolescents in a non-clinical setting. The dataset covers five views across different devices and natural illumination. We compare our method with representative contrastive and masked-image-modeling baselines on three aspects of dental-arch awareness: local-to-global localization, tooth and sextant feature probing, and cross-subject consistency with FDI tooth relations. Qualitative dense-feature visualizations further examine whether the learned representation captures the semantics of different teeth and oral tissues. Finally, we assess the transferability of our pretrained representation to tooth segmentation.

Our contributions are threefold. (1) We introduce YoungDent, a large-scale, five-view intraoral-photograph dataset captured with smartphones under diverse, non-standardized conditions, providing a practical basis for 2D dental representation learning; (2) We propose a self-supervised framework that combines dental-arch-aware local-to-global feature alignment, FDI tooth relation modeling, and masked reconstruction across tooth, sextant, and full-mouth scales; (3) We establish a comprehensive evaluation of dental representations and show that our method learns stronger dental-arch-aware features than generic SSL methods.

2 Related Work

Self-Supervised Learning SSL methods learn transferable visual representations without manual annotation. Contrastive approaches such as SimCLRv2 and MoCo enforce invariance across augmented views while separating different images (Chen et al., 2020; He et al., 2020). SwAV and SlotCon instead use clustering or semantic grouping to capture structure beyond instance discrimination (Caron et al., 2020; Wen et al., 2022). The DINO family removes explicit negative pairs and uses teacher–student self-distillation across global and local crops, producing strong semantic and dense features (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025). In parallel, masked image modeling (MIM) learns contextual representations by recovering hidden content, either as discrete tokens in BEiT or pixels in MAE (Bao et al., 2021; He et al., 2022a; Xie et al., 2022; Wei et al., 2022). Contrastive and distillation objectives emphasize semantic invariance, whereas reconstruction preserves fine-grained appearance. Hybrid methods combine their complementary strengths (Huang et al., 2023b). To better support downstream tasks requiring local descriptions, such as object detection and semantic segmentation, SSL objectives have also been extended from image-level to patch-level (Hénaff et al., 2021; Wei et al., 2021; Xie et al., 2021a;b). Of particular relevance to our work, position-based objectives use relative locations or spatial correspondences as

supervision, showing that geometric structure can support part-level representation learning (Doersch et al., 2015; Zhang & Gong, 2023).

SSL in Medical Imaging SSL has been widely applied to radiographs, MRI, and histopathology, where labeled data are expensive and specialized expertise is required (Huang et al., 2023a; Sowrirajan et al., 2021; Li et al., 2021a; Wang et al., 2021). Recent large-scale models further demonstrate the value of self-supervised pretraining across multiple medical imaging tasks (Moutakanni et al., 2026; Kim et al., 2026; Yao et al., 2025). Despite this progress, objectives designed for natural images do not explicitly model the regular spatial organization of organs and tissues. Such anatomical structure is more consistent across medical images than across natural images, making it a valuable source of domain-specific supervision. TransVW treats recurring anatomical patterns in chest X-rays as transferable visual words (Haghighi et al., 2021). RSP and SSLP introduce consistency training and spatial guidance for histology whole-slide images (Srinidhi et al., 2022; Li et al., 2021b), and VoCo predicts the overlap between a random crop and its source 3D volume (Wu et al., 2024). Together, these studies show that anatomical and geometric priors provide information beyond generic view invariance. We extend this principle to intraoral photographs, where tooth position and inter-dental-arch relations form a consistent structure across subjects.

Tooth Structure Understanding Most learning-based dental studies focus on task-specific tooth segmentation in CBCT or IOS data (Cui et al., 2021; Xiong et al., 2023; Xi et al., 2025; Krenmayr et al., 2024). Though their specialized architectures can capture detailed 3D geometry, the learned features are not designed for transfer across different dental tasks. Recent work has introduced self-supervised pretraining for 3D tooth segmentation (He et al., 2022b; Almalki & Latecki, 2024), and ToothMCL aligns CBCT and IOS representations through multimodal contrastive learning (Son et al., 2025). DentalMIM and IGLC-MAE adapt MIM-style pretraining for dental panoramic radiographs (Almalki & Latecki, 2023; Zhang et al., 2026). However, these methods still depend on specialized clinical acquisition and primarily optimize task- or modality-level representations, leaving inter-dental-arch relations and dense, structure-aware representations largely unexplored. Representation learning from more accessible intraoral photographs is also underexplored. We address this gap by linking local dense features to the global dentition, enabling representations to distinguish dental tissues and tooth regions while preserving anatomical relations specified by tooth identity and position.

3 METHOD

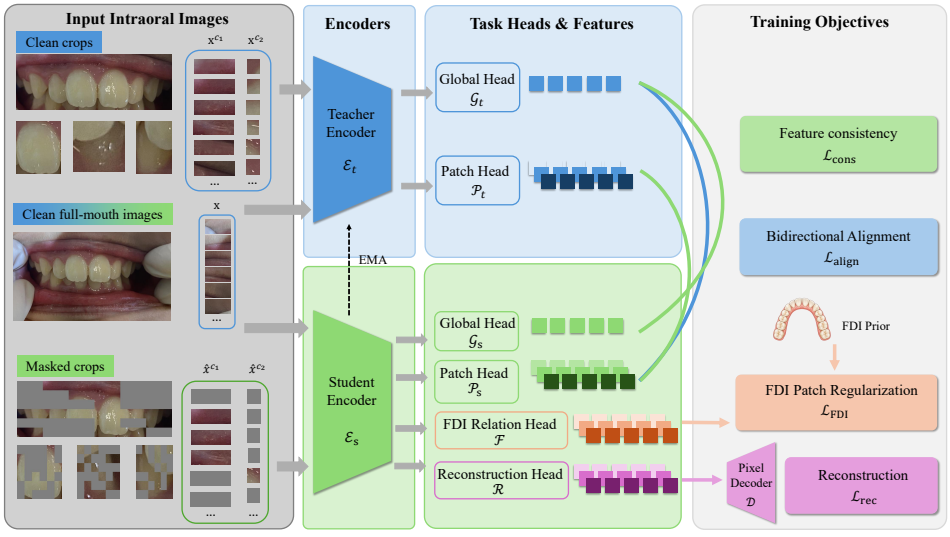


Figure 2: Overview of our method. The momentum teacher processes clean full-mouth images and crops, whereas the student receives clean full-mouth images and masked crops. Global and patch heads support bidirectional local-to-global feature alignment and masked–clean feature consistency. An FDI relation head regularizes student patch features using dental prior, while a reconstruction head and pixel decoder recover masked crop patches.

Problem formulation Given a full-mouth intraoral image I, a pretrained tooth-numbering model (*e.g.* SegmentAnyTooth (Nguyen et al., 2025)) predicts tooth bounding boxes and their FDI codes, $\mathcal{T}(I) = \{(B_k, a_k)\}_{k=1}^{K}$, where B_k and a_k are the bounding box and FDI code of tooth k. From these predictions, we construct single-tooth and sextant crops, together with multi-scale random crops. For a crop indexed by c, we denote the crop image as $I_c = \mathrm{Crop}(I, B_c)$. A single-tooth crop also has an FDI code a_c. We pursue two objectives: to align each crop with dense features at its corresponding full-image location, enforcing position-aware local-to-global consistency; and to organize tooth representations according to FDI-defined relations among tooth identities.

FDI tooth relation prior Dental anatomy induces graded geometric and morphological relations. The same tooth across subjects should be most similar, followed by contralateral and morphologically corresponding teeth. We encode these relations with an FDI-based score $R(a, b) \in [0, 1]$. As shown in Fig. 4, this prior supervises both single-tooth feature alignment and full-image patch regularization.

Overview Fig. 2 illustrates our momentum teacher–student architecture. Each crop representation is aligned with dense features at the corresponding full-image location using positional overlap as the target (Sec. 3.1). For single-tooth crops, FDI tooth relations reduce the negative penalties assigned to the representation of anatomically related, non-overlapping teeth. The same prior organizes dense full-image features (Sec. 3.2), while masked reconstruction and masked-clean consistency retain fine-grained appearance and improve robustness to partial observations (Sec. 3.3).

The teacher and student share the same backbone architecture, using either a DINO-style Vision Transformer (Oquab et al., 2023) or a Swin Transformer (Liu et al., 2021). Separate projection heads prevent the alignment, reconstruction, consistency, and FDI-relation objectives from competing in one feature space. After pretraining, only the teacher encoder is retained for downstream tasks.

3.1 Local-to-Global Feature Alignment

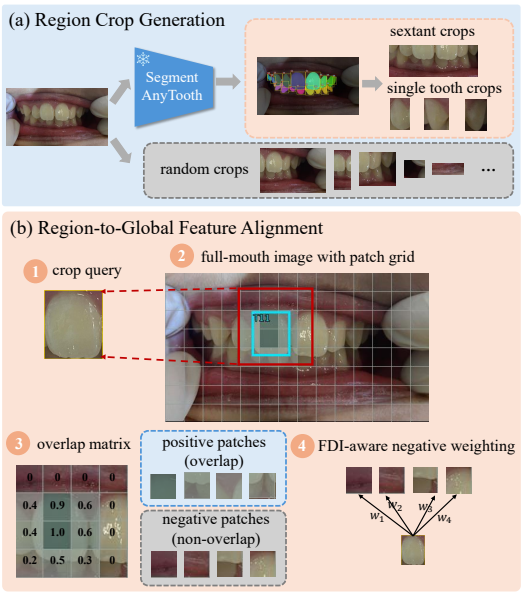


Figure 3: Dental-arch-aware local-to-global feature alignment. (a) Structural and random crops generation. (b) Align single-tooth crop with the patch grid of its source full-mouth image using their overlap matrix, while non-overlapping tooth patches receive FDI-aware negative weights according to their anatomical relation to the crop tooth.

Each crop has a known position in its source full-mouth image. As shown in Fig. 3, we match the similarity between its representation and each full-image patch to their positional overlap. The projected global student feature of crop c is $z^g_{s,c} = \mathcal{G}_s(\mathcal{E}_s(\hat{x^c}))$, where $\hat{x^c}$ is the patchified masked crop, and the projected teacher feature of full-image patch i is $z^p_{t,i} = \mathcal{P}_t(\mathcal{E}_t(x))$. Let $s_{c,i}$ denote their cosine similarity, and $y_{c,i} \in [0, 1]$ the overlap between crop c and patch i (*i.e.* the fraction of crop area covered by a patch). Their soft distance is

$$\ell(s_{c,i}, y_{c,i}) = -\log(1 - |y_{c,i} - [s_{c,i}]_+| + \epsilon)$$

where $[s]_+ = \max(s, 0)$.

FDI-Aware negative weighting Uniformly penalizing all non-overlapping patches can incorrectly separate anatomically related teeth. For a single-tooth crop c, let a_i be the FDI code of patch i and set $R(a_c, a_i) = 0$ for background patches. We define the alignment weight as

$$w_{c,i} = \begin{cases} 1, & y_{c,i} > 0, \\ 0.1 + 0.9\big(1 - R(a_c, a_i)\big), & y_{c,i} = 0 \end{cases}$$

so that background and anatomically unrelated teeth remain strong negatives, while related teeth receive weaker penalties. Because sextant and random crops have no unique FDI identity, we set $w_{c,i} = 1$ for them. For the crop set $\mathcal{C}$, the crop-to-image alignment loss is

$$\mathcal{L}^{s\to t}_{\text{align}} = \frac{1}{|\mathcal{C}|} \sum_{c\in\mathcal{C}} \frac{\sum_i w_{c,i}\, \ell(s_{c,i}, y_{c,i})}{\sum_i w_{c,i}} \tag{1}$$

Bidirectional alignment We also align the global teacher feature of each clean crop, $z^g_{t,c} = \mathcal{G}_t(\mathcal{E}_t(x_c))$, with the student patch features of the clean full image, $z^p_{s,i} = \mathcal{P}_s(\mathcal{E}_s(x))$. The reverse loss $\mathcal{L}^{t\to s}_{\text{align}}$ uses the same overlap targets and alignment weights. The complete alignment objective is $\mathcal{L}_{\text{align}} = \mathcal{L}^{s\to t}_{\text{align}} + \lambda_{bi}\mathcal{L}^{t\to s}_{\text{align}}$.

3.2 FDI-Based Patch Regularization

Local-to-global alignment constrains crop–patch correspondence but does not fully organize the full-image patch embeddings. We therefore apply the FDI tooth prior to student patch features projected by the FDI relation head, $f_i = \mathcal{F}(\mathcal{E}_s(x))$. Each full-image patch i is assigned according to its maximum overlap with a tooth box. If this overlap exceeds τ_{tooth}, the patch enters the tooth set $\mathcal{S}_t$ and inherits the corresponding FDI code a_i. Non-overlapping patches form the background set $\mathcal{S}_b$, while ambiguous patches with overlap in $(0, \tau_{\text{tooth}}]$ are excluded.

For a patch $i \in \mathcal{S}_t$, we define a predicted similarity distribution over candidate tooth patch n and an FDI-derived target distribution:

$$O_i(n) = \frac{\exp(f_i^\top f_n/\tau)}{\sum_{j\in\mathcal{S}_t\setminus\{i\}}\exp(f_i^\top f_j/\tau)},$$

$$T_i(n) = \frac{\exp(R(a_i,a_n)/\tau)/N_{a_n}}{\sum_{j\in\mathcal{S}_t\setminus\{i\}}\exp(R(a_i,a_j)/\tau)/N_{a_j}}$$

where N_a is the number of candidate patches assigned to FDI code a. The tooth-to-tooth regularizer $\mathcal{L}^{\text{tt}}_{\text{FDI}}$ in Eq. 2 preserves graded anatomical relations by KL divergence without imposing an absolute similarity on each tooth pair. To separate tooth features from heterogeneous background content, we also penalize positive tooth-to-background similarities $\mathcal{L}^{\text{tb}}_{\text{FDI}}$. Background-to-background pairs remain unconstrained because they have no defined anatomical relation.

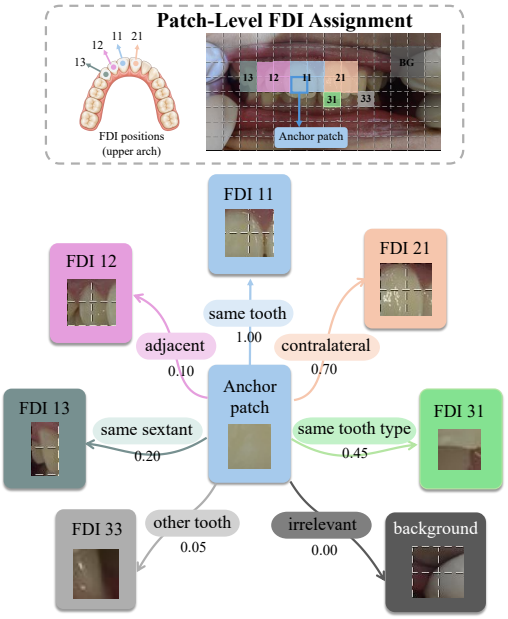


Figure 4: FDI tooth relation prior. Using an anchor patch from tooth 11 as an example, we assign other patches graded relation scores according to FDI anatomy: the same tooth, contralateral tooth, same tooth category, same sextant, adjacent tooth, other teeth, and background. In practice, each tooth-containing patch serves as an anchor for FDI-relation modeling.

$$\mathcal{L}^{\text{tt}}_{\text{FDI}} = \frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t} D_{\text{KL}}(T_i \,\|\, O_i) \qquad \mathcal{L}^{\text{tb}}_{\text{FDI}} = \frac{1}{|\mathcal{S}_t||\mathcal{S}_b|}\sum_{i\in\mathcal{S}_t}\sum_{n\in\mathcal{S}_b}[f_i^\top f_n]^2_+ \tag{2}$$

The complete patch regularizer is $\mathcal{L}_{\text{FDI}} = \mathcal{L}^{\text{tt}}_{\text{FDI}} + \mathcal{L}^{\text{tb}}_{\text{FDI}}$.

3.3 Masked Representation Learning and Overall Objective

Masked image modeling Following DINOv2 (Oquab et al., 2023), the masked–clean consistency loss $\mathcal{L}_{\text{cons}}$ aligns the student's masked-crop representation with the teacher's clean-crop representation. An MAE-style reconstruction loss $\mathcal{L}_{\text{rec}}$ (He et al., 2022a) further recovers masked patches from student features, preserving appearance such as crown morphology that alignment may overlook.

Overall objective The final objective combines local-to-global alignment, masked reconstruction, masked-clean consistency, and FDI-based patch regularization.

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{FDI}}\mathcal{L}_{\text{FDI}}. \tag{3}$$

4 Experiments

4.1 Experimental Setup

Datasets We use three intraoral-photograph datasets: our YoungDent dataset and the public AIDC (Faizan Ahmed et al., 2025) and Odontify (Aranguiz & Tobar, 2026) datasets. We use SegmentAnyTooth (Nguyen et al., 2025) to predict tooth positions and FDI coding, from which we obtain over 50K tooth crops and 6K sextant crops within YoungDent. We convert AIDC to the same format and combine it with YoungDent using an 80/20 subject-level split for pretraining and evaluation. Odontify is used only for out-of-domain (OOD) tooth segmentation evaluation.

Table 1: Dataset comparison. YoungDent is collected in this work. AIDC and Odontify are public datasets. *5-view ratio* is the percentage of subjects with all five intraoral views.

Metric		YoungDent	AIDC	Odontify
# Subjects		1027	1,113	179
# Images		5,377	6,265	891
5-view ratio		98.1%	57.3%	98.9%
Resolution	Full	2599×1961	1635×1178	572×391
	Sextant	1175×517	1029×424	423×174
	Tooth	244×238	231×224	68×75

Baselines & Implementation We compare against the masked image modeling method SwinMAE (He et al., 2022a; Xu et al., 2023) and the contrastive learning methods SimCLRv2 (Chen et al., 2020) and DINOv2 (Oquab et al., 2023). Evaluation uses the unprojected output of the final teacher layer for DINOv2 and our method, and the final encoder output for the other baselines.

Our Swin model is initialized from Swin-S pretrained on ImageNet-22K (Deng et al., 2009), while the ViT model uses DINOv2 ViT-S/14 initialization. Pretraining proceeds through low- and high-resolution stages on NVIDIA 3090 GPUs, using the AdamW optimizer (Loshchilov & Hutter, 2017), peak learning rates of 1×10^{-4} and 5×10^{-5}, respectively, and a cosine learning-rate schedule. Each sample contains one full-mouth image, 8 tooth crops, 2 sextant crops, and 8 random crops. By default, we show the results of our Swin model. App. B provides the remaining settings.

4.2 Comparison on Dental-Arch Awareness

Evaluation settings We assess dental-arch awareness using three quantitative tasks: *Localization*, which retrieves the source region of a crop from full-image patch similarities; *Feature probing on teeth and sextants*, which evaluates frozen crop features with KNN and logistic regression; and *Cross-subject FDI tooth relation*, which tests whether tooth representations preserve anatomical relations across subjects. PCA visualizes dense semantic structure and local-to-global consistency, while UMAP (McInnes et al., 2018) visualizes tooth and sextant representation spaces.

Table 2: Localization and feature probing ability across different SSL methods. *random crop* and *structure crop* denote that we use only random crops or single-tooth and sextant crops, respectively. Macro-F1 is reported for different probing tasks.

		Localization					Probing (KNN / Logistic)		
Method	**Backbone**	**GT Sim. ↑**	**Out. Sim. ↓**	**Sim. Con. ↑**	**Patch R@1/10 ↑**	**IoU ↑**	**Tooth FDI ↑**	**Tooth Class ↑**	**Sextant ↑**
Masked Autoencoding									
Swin-MAE	Swin-S	0.922	0.909	0.013	0.080 / 0.349	0.047	0.219 / 0.289	0.404 / 0.445	0.655 / 0.775
Contrastive Learning									
SimCLRv2	ResNet50	0.224	0.151	0.073	0.300 / 0.716	0.153	0.707 / 0.714	0.822 / 0.822	0.935 / 0.937
DINOv2	ViT-S/14	0.585	0.414	0.171	0.343 / 0.793	0.221	0.346 / 0.596	0.596 / 0.707	0.882 / 0.932
Masked Alignment Pretraining									
Ours (random crop)	Swin-T	0.743	0.657	0.086	0.332 / 0.613	0.165	0.633 / 0.748	0.701 / 0.762	0.949 / **0.962**
Ours (structural crop)	Swin-T	0.670	0.625	0.045	0.297 / 0.619	0.107	0.592 / 0.777	0.827 / 0.837	0.925 / 0.952
Ours	Swin-T	0.736	0.647	0.089	0.508 / 0.777	0.245	0.656 / 0.759	0.827 / 0.830	0.936 / 0.937
Ours	Swin-S	0.663	0.365	**0.298**	**0.866 / 0.985**	**0.485**	**0.747 / 0.793**	**0.834 / 0.844**	**0.962** / 0.946
Ours	ViT-S/14	0.815	0.645	0.170	0.506 / 0.759	0.287	0.403 / 0.541	0.509 / 0.626	0.907 / 0.924

Localization For crop c, we compare its global teacher feature $\tilde{z}^g_{t,c}$ with the full-image patch features $\{\tilde{z}^p_{t,i}\}_{i=1}^N$, where $\tilde{z}$ denotes an unprojected backbone feature. *GT Similarity* and *Outside Similarity* average the cosine similarity over overlapping and non-overlapping patches, and their difference gives *Similarity Contrast*. *Patch Recall@*k ($k \in \{1, 10\}$) records whether an overlapping patch appears among the top-k valid patches. We also slide a window of the ground-truth crop size over the similarity map and report the *IoU* between the highest-scoring window and the target box. App. C.1 provides full definitions.

Tab. 2 shows that our Swin-S model achieves the strongest box-level localization. Compared with DINOv2, it improves Patch R@1 from 0.343 to 0.866 and IoU from 0.221 to 0.485. It also attains the highest similarity contrast (0.298), indicating more discriminative features. In contrast, the near-zero score of Swin-MAE (0.013) suggests its limited discrimination of dental-arch semantics. Fig. 5

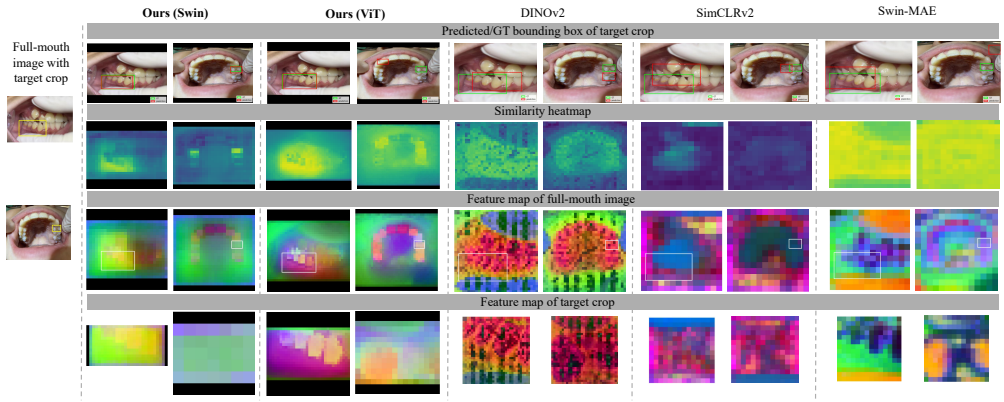


Figure 5: Qualitative comparison of structural crop localization and local-to-global feature consistency. For a target crop, we show predicted and ground-truth boxes, crop-to-image patch similarity heatmap, and PCA feature maps. Across Swin and ViT backbones, our method concentrates similarity on the target, preserves discriminative tooth semantics, and aligns crop features with their full-image counterparts.

provides qualitative support: the similarity response of our model concentrates within the target crop and yields a box closer to the ground-truth location than those of generic SSL baselines.

Feature probing on teeth and sextants We freeze each encoder and train cosine KNN and logistic regression probes for 32 FDI codes, eight tooth categories, and six sextant types. In Tab. 2, our model outperforms Swin-MAE, DINOv2, and SimCLRv2 on every probing task. Our Swin-S model achieves KNN F1 scores of 0.834 and 0.962 for tooth-category and sextant classification, respectively, surpassing the strongest SimCLRv2 results. Logistic regression shows the same trend.

Notably, the crop variants reveal complementary supervision. Random crops favor regional context and localization, outperforming structural crops in Patch R@1 (0.332 versus 0.297) and IoU (0.165 versus 0.107). Structural crops better encode individual-tooth semantics, achieving logistic-probe F1 scores of 0.777 for FDI codes and 0.837 for tooth categories. Combining both yields the best localization (0.508 Patch R@1 and 0.245 IoU) while maintaining balanced tooth- and sextant-level performance. Scaling the backbone from Swin-T to Swin-S further improves localization (IoU: $0.245 \rightarrow 0.485$) and probing performance (FDI KNN: $0.656 \rightarrow 0.747$). Our ViT variant performs worse than Swin in both tasks, suggesting that the windowed attention and multi-scale design of Swin are better suited to dental-arch-aware representation learning under our SSL objectives.

Table 3: Cross-subject similarity by FDI tooth relation. Entries are mean cosine similarities between cross-subject single-tooth pairs with standard deviation.

FDI Relation	Swin-MAE	SimCLRv2	DINOv2	Ours (Swin)	Ours (ViT)
Weak relation	$0.913_{\pm 0.09}$	$0.299_{\pm 0.09}$	$0.714_{\pm 0.13}$	$0.675_{\pm 0.10}$	$0.597_{\pm 0.25}$
Same sextant	$0.932_{\pm 0.07}$	$0.360_{\pm 0.12}$	$0.737_{\pm 0.12}$	$0.725_{\pm 0.11}$	$0.696_{\pm 0.22}$
Same category	$0.929_{\pm 0.08}$	$0.407_{\pm 0.12}$	$0.741_{\pm 0.12}$	$0.757_{\pm 0.09}$	$0.701_{\pm 0.18}$
Adjacent	$0.935_{\pm 0.07}$	$0.516_{\pm 0.16}$	$0.760_{\pm 0.11}$	$0.844_{\pm 0.07}$	$0.789_{\pm 0.18}$
Symmetric	$0.938_{\pm 0.07}$	$0.602_{\pm 0.17}$	$0.772_{\pm 0.11}$	$0.911_{\pm 0.06}$	$0.813_{\pm 0.17}$
Same FDI	$0.938_{\pm 0.07}$	$0.782_{\pm 0.15}$	$0.774_{\pm 0.11}$	$0.915_{\pm 0.06}$	$0.821_{\pm 0.17}$
Overall Spearman	0.232	0.651	0.301	**0.680**	0.522

Cross-Subject FDI relation For single-tooth crop pairs from different test subjects, we compare cosine similarity partitioned by FDI relation defined in Fig. 4. We report the mean similarity for each relation and the Spearman correlation between FDI relation scores and the aggregated 32×32 mean-similarity matrix (Fig. 6(a)).

In Tab. 3, similarity increases with anatomical relation strength. Our methods provide clearer ranking pattern over Swin-MAE and DINOv2, achieving the highest Spearman correlation with predefined FDI relation of 0.68.

Consistently, Fig. 6(a) shows a pronounced diagonal and high similarity between contralateral teeth of the same morphology (*e.g.* $(11, 21)$, $(17, 27)$, and $(31, 41)$).

Visualization of representation Fig. 5 provides further qualitative evidence. With either Swin or ViT, our method assigns distinct yet anatomically coherent feature colors to different teeth in the full-image maps. Contralateral teeth with the same FDI identity show similar features. It also yields clearer local-to-global consistency, where crop features retain similar color family as their corresponding full-image regions while resolving progressively finer details. This pattern is not

clearly observed in single-tooth crop feature map from SimCLRv2 and Swin-MAE. We observe a potential trade-off: by encouraging similar representations for anatomically related teeth, FDI regularization may reduce localization specificity when visually similar teeth compete for the same query.

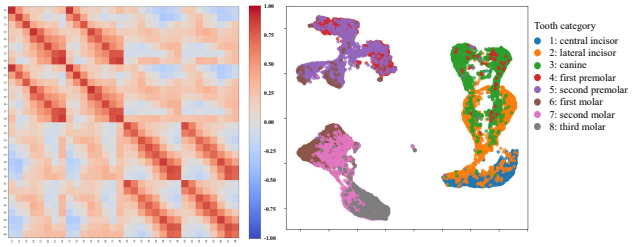


Figure 6: Anatomical structure in single-tooth representations. (a) Cross-subject similarity matrix indexed by FDI code. (b) UMAP of upper-occlusal tooth-crop features, colored by tooth category.

The sextant UMAP in Fig. 7 forms compact groups for distinct regions while keeping symmetric sextants nearby (*e.g.*, S1/S3 and S4/S6). This structure is clearer than that of Swin-MAE and better balances separation and anatomical similarity than DINOv2 and SimCLRv2. Likewise, the tooth UMAP in Fig. 6(b) groups samples by morphology, confirming that the feature space captures fine-grained dental semantics. App. D.1 provides more results in qualitative and quantitative comparisons.

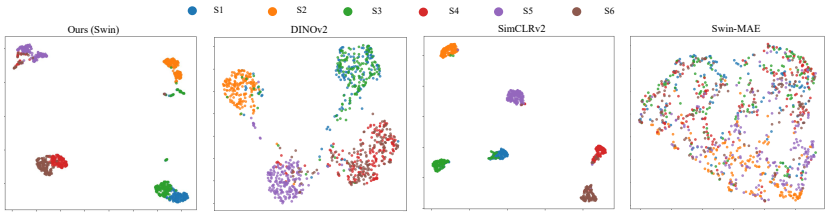


Figure 7: UMAP visualization of sextant-crop representations. Points are colored by sextant label. Our method forms compact, well-separated sextant groups while retaining proximity between anatomically symmetric sextants (*e.g.*, S1/S3 and S4/S6).

4.3 Ablation Study

Table 4: Ablation of crop configuration and training objectives. Probing reports logistic-regression macro-F1. *Relation range* denotes the max−min gap between the mean cross-subject tooth similarities of the FDI relation categories. *# Struc crop* denotes the number of single-tooth / sextant crops, and *large res* uses higher input resolution. Green and red backgrounds indicate the best and worst values, respectively, in each metric column.

						localization		probing			FDI relation
#struc crop	#rand crop	large res	L_{FDI}	L_{cons}	L_{rec}	patch@1	IoU	sextant	tooth class	tooth FDI	relation range
6/2	4	✗	✓	✓	✓	0.607	0.263	0.951	0.839	0.780	0.227
12/2	4	✗	✓	✓	✓	0.663	0.295	0.962	0.838	0.786	0.279
6/2	8	✗	✓	✓	✓	0.647	0.286	0.950	0.846	0.770	0.165
6/2	4	✓	✓	✓	✓	0.698	0.382	0.948	0.837	0.775	0.228
6/2	4	✗	✗	✓	✓	0.621	0.299	0.944	0.827	0.750	0.173
6/2	4	✗	✓	✗	✓	0.630	0.297	0.954	0.837	0.764	0.230
6/2	4	✗	✓	✓	✗	0.593	0.288	0.956	0.851	0.739	0.542

Effect of crop number and resolution In Tab. 4, more structural crops provide clear localization gains (Patch@1 $0.607 \rightarrow 0.663$) and tooth semantic probing advantages (sextant $0.951 \rightarrow 0.962$). Increasing random crops instead improves tooth-category probing ($0.839 \rightarrow 0.846$) but reduces the FDI relation range, reflecting a trade-off between broad regional context and tooth identity. Higher-resolution training achieves the strongest localization ability.

Effect of loss functions Removing $\mathcal{L}_{\text{FDI}}$ decreases tooth-category probing, FDI probing, and the relation range, confirming that it encodes anatomical tooth relations. Removing $\mathcal{L}_{\text{rec}}$ produces the lowest Patch@1 and FDI probing score, despite the largest relation range. This larger range does not imply better FDI modeling. Without reconstruction, the FDI and alignment objectives may dominate the representation and exaggerate this range while weakening localization and tooth semantics. Reconstruction anchors the features to fine-grained visual evidence, yielding a more balanced representation. Removing $\mathcal{L}_{\text{cons}}$ retains competitive localization but reduces FDI probing, indicating its auxiliary contribution to dental semantics.

4.4 Tooth Segmentation

Evaluation setup For each SSL method, we freeze the pretrained encoder and train either a linear head or a decoder for 33-class segmentation (32 FDI classes and background) on 448×448

images. Both heads are trained with 10% or 100% of labeled YoungDent and evaluated in-domain. DeepLabV3 (Chen et al., 2017) and Swin-UPerNet (Liu et al., 2021) serve as end-to-end references without SSL pretraining. All models are also evaluated on Odontify (Aranguiz & Tobar, 2026) without retraining to assess OOD transfer. We report mIoU over eight tooth-type groups (T_1–T_8). App. C.3 details the settings.

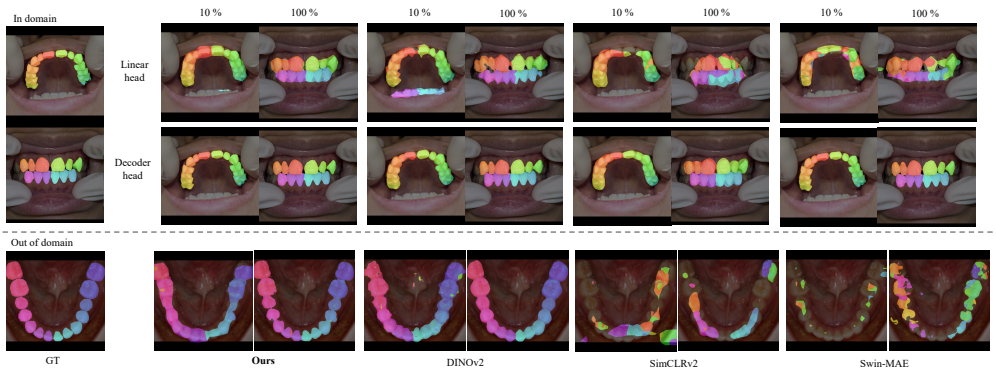


Figure 8: Qualitative tooth segmentation results using frozen SSL encoders with segmentation heads. We compare linear heads and decoders trained with 10% or 100% of the YoungDent training set, evaluated in-domain on YoungDent and out-of-domain on Odontify (Aranguiz & Tobar, 2026).

Table 5: Tooth segmentation mIoU under in-domain and out-of-domain settings. Frozen pretrained encoders with linear/decoder heads and end-to-end methods are evaluated with 10% and 100% labeled training data.

		In-domain				Out-of-domain			
		10%		100%		10%		100%	
Setting	**Method**	Linear	Decoder	Linear	Decoder	Linear	Decoder	Linear	Decoder
Pretraining +Head	Swin-MAE	0.094	0.535	0.119	0.614	0.029	0.224	0.036	0.317
	SimCLRv2	0.220	0.396	0.256	0.448	0.104	0.233	0.125	0.278
	DINOv2	0.257	0.510	0.291	0.574	0.169	0.310	0.197	0.371
	Ours	0.392	**0.634**	0.411	<u>0.671</u>	<u>0.272</u>	**0.367**	0.283	**0.431**
End-to-End	DeepLabV3	<u>0.621</u>		**0.720**		0.199		<u>0.309</u>	
	Swin-UPerNet	0.510		0.671		0.209		0.221	

Segmentation across heads, training set sizes, and domains Among frozen SSL encoders, our method achieves the highest mIoU for every head, data scale, and domain setting (Tab. 5). With 10% of the training data, it obtains $0.392/0.634$ mIoU with the linear head/decoder in-domain and $0.272/0.367$ OOD. Full supervision further raises the performance. The strong linear-head results indicate that the frozen encoder learns transferable dense features, while the decoder further exploits multi-scale structure.

Fig. 8 corroborates these results. The linear head recovers coarse tooth identities, and the decoder refines individual boundaries. With 10% labels, our decoder exceeds DeepLabV3 and Swin-UPerNet in-domain. With full labels, it matches Swin-UPerNet and approaches DeepLabV3. It also consistently surpasses both end-to-end models in the OOD setting with the decoder head, suggesting stronger transferability of the tooth-aware representation. App. D.2 provides more quantitative results on per-category segmentation and visualizations.

5 Conclusion

We present a masked alignment pretraining framework for learning representations from intraoral photographs. It aligns random, single-tooth, and sextant crops with dense features at their corresponding locations in full-mouth images, and use FDI tooth relation to organize tooth features anatomically. We further contribute YoungDent, an intraoral image dataset acquired in non-clinical settings. Compared with generic SSL methods, our method learns clearer local-to-global correspondence and more anatomically organized dental-arch features. The learned representations also transfer effectively to tooth segmentation, achieving competitive performance against end-to-end models. These results establish dental arch geometry and FDI tooth relation as effective priors, providing a practical foundation for transferable dental image representation learning.

AI USE STATEMENT

In this work, we use generative AI tools to polish the text, refine the pipeline figure, summarize relevant literature, and assist with engineering implementation. We have not used generative AI tools for core methodology design, critical mathematical derivations, or primary experiment design. We have thoroughly reviewed, verified, and edited all AI-assisted content for factual accuracy, consistency, and correctness. We take full responsibility for the final content of this work, including all written text and scientific claims.

ETHICS STATEMENT

This work strictly adheres to high standards of research integrity and ethical guidelines. For the dataset collected in this study (*YoungDent*), intraoral photograph acquisition from student participants received approval from the Institutional Review Board (IRB). Written informed consent was explicitly obtained from the school authorities, students, and their parent or legal guardian prior to data collection. The entire process was conducted in accordance with the principles of voluntary participation and privacy protection. All collected images were anonymized by strictly removing facial identity information unrelated to dental structures. For other public datasets used in our model training, we strictly comply with their respective terms of use, utilizing them solely for academic research purposes. Furthermore, this study poses no potential negative social impacts.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility of our work, we provide comprehensive implementation and experimental details in the Appendix. Specifically, we elaborate on the hyperparameter configurations, network architectures, and dataset preprocessing protocols. In addition, the Appendix details the evaluation methodologies used for all baseline comparisons as well as the setup for extended comparative and ablation studies. Upon acceptance of the paper, we will release the complete codebase, data processing scripts, and pretrained model weights to the research community to support further research.

REFERENCES

Amani Almalki and Longin Jan Latecki. Self-supervised learning with masked image modeling for teeth numbering, detection of dental restorations, and instance segmentation in dental panoramic radiographs. In *WACV*, 2023.

Amani Almalki and Longin Jan Latecki. Self-supervised learning with masked autoencoders for teeth segmentation from intra-oral 3d scans. In *WACV*, 2024.

Leonardo Aranguiz and Eduardo Tobar. Segmentation teeth images and masks, 2026. URL `https://www.kaggle.com/dsv/16143375`.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *ICLR*, 2026.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inform. Process. Syst.*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Adv. Neural Inform. Process. Syst.*, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021.

Zhiming Cui, Changjian Li, Nenglun Chen, Guodong Wei, Runnan Chen, Yuanfeng Zhou, Dinggang Shen, and Wenping Wang. Tsegnet: An efficient and accurate tooth segmentation network on 3d dental model. *Medical image analysis*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

Syed Muhammad Faizan Ahmed, Muhammad Huzaifa Ghori, Aamna Khalid, Ayesha Nooruddin, Niha Adnan, Abhishek Lal, and Fahad Umer. Annotated intraoral image dataset for dental caries detection. *Scientific data*, 2025.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE transactions on medical imaging*, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022a.

Xiaoxuan He, Hualiang Wang, Haoji Hu, Jianfei Yang, Yang Feng, Gaoang Wang, and Liu Zuozhu. Unsupervised pre-training improves tooth segmentation in 3-dimensional intraoral mesh scans. In *International Conference on Medical Imaging with Deep Learning*, 2022b.

Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron Van den Oord, Oriol Vinyals, and Joao Carreira. Efficient visual pretraining with contrastive detection. In *ICCV*, 2021.

Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P Lungren, Serena Yeung, and Akshay S Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ digital medicine*, 2023a.

Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE TPAMI*, 2023b.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. In *Adv. Neural Inform. Process. Syst.*, 2023.

Shinn Kim, Soobin Lee, Kyoungseob Shin, Han-Soo Kim, Yongsung Kim, Minsu Kim, Juhong Nam, Somang Ko, Daeheon Kwon, Wook Huh, et al. A large-scale vision foundation model for musculoskeletal radiographs. *npj Digital Medicine*, 2026.

Lucas Krenmayr, Reinhold von Schwerin, Daniel Schaudt, Pascal Riedel, and Alexander Hafner. Dilatedtoothsegnet: Tooth segmentation network on 3d dental meshes through increasing receptive vision. *Journal of Imaging Informatics in Medicine*, 2024.

Hongwei Li, Fei-Fei Xue, Krishna Chaitanya, Shengda Luo, Ivan Ezhov, Benedikt Wiestler, Jianguo Zhang, and Bjoern Menze. Imbalance-aware self-supervised learning for 3d radiomic representations. In *MICCAI*, 2021a.

Jiajun Li, Tiancheng Lin, and Yi Xu. Sslp: Spatial guided self-supervised learning on pathological images. In *MICCAI*, 2021b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Yi Liu, Yuxi Cheng, Yang Song, Daoheng Cai, and Niankun Zhang. Oral screening of dental calculus, gingivitis and dental caries through segmentation on intraoral photographic images using deep learning. *BMC Oral Health*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Théo Moutakanni, Piotr Bojanowski, Guillaume Chassagnon, Céline Hudelot, Armand Joulin, Yann LeCun, Matthew Muckley, Maxime Oquab, Marie-Pierre Revel, and Maria Vakalopoulou. Advancing human-centric ai for robust x-ray analysis through holistic self-supervised learning. *Nature Communications*, 2026.

Khoa D. Nguyen, Huy T. Hoang, Thi-Phuong-Hoa Doan, Kim-Quyen Dao, Ding-Han Wang, and Min-Ling Hsu. Segmentanytooth: An open-source deep learning framework for tooth enumeration and segmentation in intraoral photos. *Journal of Dental Sciences*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Moo Hyun Son, Juyoung Bae, Zelin Qiu, Jiale Peng, Kai Xin Li, Yifan Lin, and Hao Chen. Multimodal contrastive pretraining of cbct and ios for enhanced tooth segmentation. *arXiv preprint arXiv:2509.07923*, 2025.

Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, 2021.

Chetan L Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L Martel. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Medical image analysis*, 2022.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *MICCAI*, 2021.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022.

Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Adv. Neural Inform. Process. Syst.*, 2021.

Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. *Adv. Neural Inform. Process. Syst.*, 2022.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *CVPR*, 2024.

Shufan Xi, Zexian Liu, Junlin Chang, Hongyu Wu, Xiaogang Wang, and Aimin Hao. 3d dental model segmentation with geometrical boundary preserving. In *CVPR*, 2025.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *ICCV*, 2021a.

Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. *Adv. Neural Inform. Process. Syst.*, 2021b.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022.

Huimin Xiong, Kunle Li, Kaiyuan Tan, Yang Feng, Joey Tianyi Zhou, Jin Hao, Haochao Ying, Jian Wu, and Zuozhu Liu. Tsegformer: 3d tooth segmentation in intraoral scans with geometry guided transformer. In *MICCAI*, 2023.

Zi'an Xu, Yin Dai, Fayu Liu, Weibing Chen, Yue Liu, Lifu Shi, Sheng Liu, and Yuhang Zhou. Swin mae: Masked autoencoders for small datasets. *Computers in biology and medicine*, 2023.

Jingfeng Yao, Xinggang Wang, Yuehao Song, Huangxuan Zhao, Jun Ma, Yajie Chen, Wenyu Liu, and Bo Wang. Eva-x: A foundation model for general chest x-ray analysis with self-supervised learning. *npj Digital Medicine*, 2025.

Qianli Zhang, Meiyu Hu, Mu Yuan, Pingyi Jia, Huajie Yu, Feng Chen, Xu-Cheng Yin, and Junran Peng. A paradigm of hybrid-supervision for annotation-scarce periapical film analysis. *BMC Oral Health*, 2026.

Zhemin Zhang and Xun Gong. Positional label for self-supervised vision transformer. In *AAAI*, 2023.

Bo Zou, Shaofeng Wang, Hao Liu, Gaoyue Sun, Yajie Wang, FeiFei Zuo, Chengbin Quan, and Youjian Zhao. Teeth-seg: An efficient instance segmentation framework for orthodontic treatment based on multi-scale aggregation and anthropic prior knowledge. In *CVPR*, 2024.

APPENDIX

This appendix is organized as follows.

App. A Datasets describes the YoungDent dataset and the public datasets used for training and evaluation.

App. B Implementation Details illustrates the architecture, the training objectives and the training configurations.

App. C Evaluation Settings defines the evaluation settings. It covers quantitative measurements on the dental-arch awareness of different pretraining methods, including the definition of localization metrics, feature probing settings using KNN and logistic regression, and cross-subject FDI relation settings. Additionally, the tooth segmentation settings are described, including the training dataset, design of segmentation heads, and end-to-end segmentation models we use.

App. D Additional Experiment Results presents additional experimental results, including how the alignment-related projection heads shape the representation, how localization ability depends on the crop type, per-class tooth segmentation results, and more visualizations on feature map and representation distribution to examine the learned dental arch semantics.

App. E Limitations and Broader Impacts discusses potential limitations and social impact.

A DATASETS

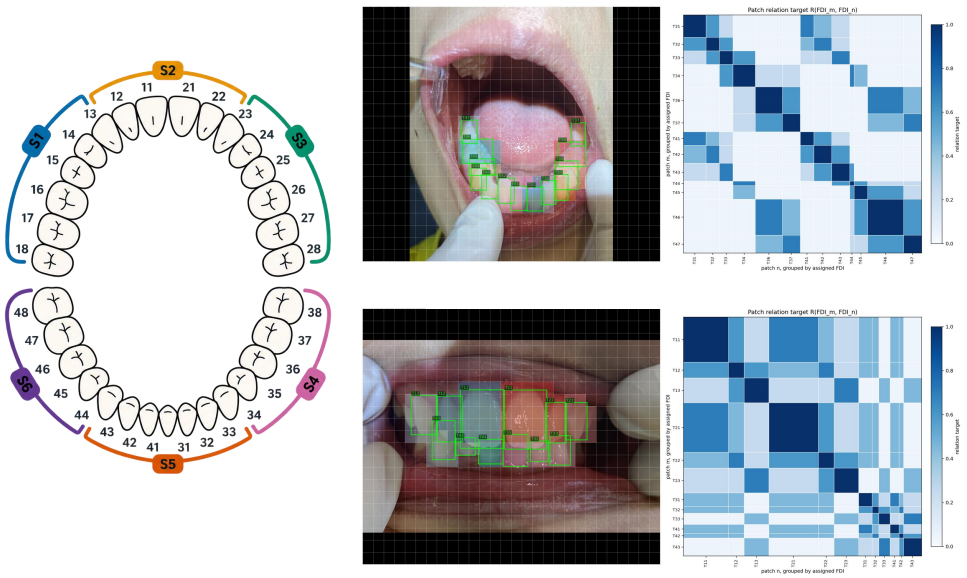
YoungDent YoungDent contains intraoral photographs of adolescents from multiple secondary schools, captured by different brands of smartphones in five views: frontal (F), left buccal (L), right buccal (R), upper occlusal (U) and lower occlusal (D). The dataset comprises 1027 subjects and 5377 intraoral images. After collection, each image is processed by SegmentAnything3 (Carion et al., 2026) to crop into a mouth-centered image (*i.e.* full-mouth images), eliminating irrelevant regions that could reveal personal identity. Then the full-mouth image is further processed by a pretrained FDI tooth number prediction model SegmentAnyTooth (Nguyen et al., 2025) to get the bounding boxes and FDI number of each single tooth within the full-mouth image. The bounding boxes of single teeth are gathered based on the predicted FDI number to obtain sextant crops. Specifically, six sextants are defined following the standard clinical division as shown in Fig. A1(a). Due to the capture range of smartphones, the frontal view exposes S_2 and S_5, the right buccal view exposes S_1 and S_6, the left buccal view exposes S_3 and S_4, and the occlusal views (U/D) expose no sextant.

Public datasets AIDC (Faizan Ahmed et al., 2025) is a public dataset containing $6,265$ images with corresponding annotations in YOLO formats for object detection of dental decays in primary and secondary dentition. It contains five-view intraoral photographs captured with and without retractors. We convert the raw images into our annotation layout, where tooth and sextant boxes are produced by SegmentAnyTooth (Nguyen et al., 2025) similar to us.

Odontify (Aranguiz & Tobar, 2026) is a public tooth-segmentation dataset from Kaggle comprising 179 patients and 891 photographs in five views. Masks are provided as per-pixel colors in the World Dental Federation notation system. This dataset is used only as an out-of-domain test set for tooth segmentation.

B IMPLEMENTATION DETAILS

FDI tooth relation prior The FDI relation prior encodes the anatomical structure of the dental arch as a graded similarity target. As illustrated in Fig. A1(a), the 32 permanent teeth are indexed using FDI notation and grouped into six sextants, S_1–S_6. For two tooth identities a and b, we assign $R(a, b) = 1.00$ when they denote the same tooth, 0.70 for contralateral teeth of the same category, 0.60 for adjacent teeth, 0.45 for teeth of the same category, and 0.25 for teeth in the same sextant. Unrelated tooth pairs receive 0.05, while background patches receive 0. Fig. A1(b) shows how this prior is transferred to image patches. Each patch is assigned the FDI code of the tooth box with which it has the largest overlap, and the pairwise relation scores are then arranged as a target matrix. The diagonal therefore has the strongest response, while contralateral and morphologically corresponding



(a) Dental arch with FDI coding

(b) FDI relation score for patches containing teeth

Figure A1: FDI relation prior. (a) FDI numbering and the six clinical sextants. (b) Patch-level relation matrices derived from FDI codes for representative images; scores are averaged over patches assigned to each tooth.

teeth form weaker but consistent off-diagonal bands. The same anatomical pattern is visible in both the frontal and occlusal examples, providing a structured target for FDI-aware alignment and patch regularization.

Architecture When a Swin Transformer is used as the encoder backbone, student and teacher share a Swin-S structure pretrained on ImageNet-22K (Deng et al., 2009). When a ViT is used as the encoder backbone, student and teacher share a ViT-S/14 structure pretrained by DINOv2 (Oquab et al., 2023). All embeddings used in cosine-similarity related objectives are ℓ_2-normalized after different projection heads.

Training objectives Given the crop-level features $(z^g_{s,c}, z^g_{t,c})$ and patch features $(z^p_{s,c,i}, z^p_{t,c,i})$, let $\omega_{c,i} \in [0,1]$ be the masked fraction of crop patch i. The masked-Clean Feature consistency loss is

$$\mathcal{L}_{\text{cons}} = \frac{1}{2|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left[1 - (z^g_{s,c})^\top z^g_{t,c} + \frac{\sum_i \omega_{c,i} \left(1 - (z^p_{s,c,i})^\top z^p_{t,c,i}\right)}{\sum_i \omega_{c,i} + \epsilon} \right] \tag{A1}$$

In the training objective defined in Eq. 3, we use $\lambda_{\text{bi}} = 0.2$, $\lambda_{\text{rec}} = 0.5$, and $\lambda_{\text{cons}} = 0.1$. To avoid imposing the FDI regularization before the establishment of stable features by alignment and reconstruction objectives, the patch-regularization weight is linearly warmed up to 0.02:

$$\lambda_{\text{patch}}(t) = 0.02 \min\left(1, \frac{t+1}{T_{\text{ramp}}}\right)$$

where t is the optimization step and T_{ramp} corresponds to the first five training epochs.

Training configurations We use a two-stage training strategy, where the first stage is at a smaller resolution before a high-resolution fine-tuning stage. In order to maintain the tooth structure within the intraoral images, letterbox resizing is applied instead of resizing. The primary configurations of two training stages are summarized in Tab. A1.

During training, we compare three different crop strategies as shown in Tab. 2: (i) *structured*: 6 single tooth crops and 2 sextant crops; (ii) *random*: 8 random crops; (iii) *hybrid* (*i.e.* default setting): 8 single tooth crops, 2 sextant crops, and 8 random crops.

Table A1: Two-stage training configuration. Stage 1 pretrains at low resolution for 100 epochs; Stage 2 adapts the Stage 1 checkpoint at high resolution for 50 epochs.

	Stage 1: low-res		Stage 2: high-res	
Backbone				
Encoder	ViT-S/14	Swin-S	ViT-S/14	Swin-S
Initialization	DINOv2	IN-22K	Stage 1	Stage 1
Resolution				
Full-image canvas	448^2	448^2	518^2	896^2
Feature grid	32×32	14×14	37×37	28×28
Single-tooth canvas	224^2	224^2	224^2	224^2
Sextant canvas	448×224	448×224	448^2	896×448
Random canvas	224^2	224^2	448^2	448^2
Masking				
Mask ratio	0.40–0.60			
Objectives				
λ_{align}	$\lambda_{\text{align}}^{\text{tooth}} = 1.0$, $\lambda_{\text{align}}^{\text{sextant}} = \lambda_{\text{align}}^{\text{random}} = 0.5$, $\lambda_{\text{bi}} = 0.2$			
λ_{rec}	0.5			
λ_{cons}	0.1			
λ_{FDI}	0.02			
Optimization				
Learning rate	$1 \times 10^{-4} \rightarrow 1 \times 10^{-5}$		$5 \times 10^{-5} \rightarrow 1 \times 10^{-6}$	
Batch size	48	16	24	8
Epochs	100	100	50	50

C Evaluation Settings

C.1 Evaluation metrics for localization

GT/Outside similarity & similarity contrast For each crop query $c \in \mathcal{C}$, let z_c^g be its ℓ_2-normalized global representation and let z_i^p be the ℓ_2-normalized representation of patch i in the corresponding full-mouth image. Consistent with Sec. 3.1, their cosine similarity and positional overlap are

$$s_{c,i} = (z_c^g)^\top z_i^p, \qquad y_{c,i} = \text{Overlap}(B_c, P_i) \in [0, 1],$$

where B_c is the ground-truth crop box and P_i is the spatial support of patch i in the full image. During calculation, we treat patch i as a valid patch when it lies inside the valid image region rather than in the letterbox padding. A crop is valid when at least 80% of it lies inside the model's valid input region. This prevents letterbox padding from creating spurious matches.

GT Similarity s_c^{GT} is the overlap-weighted mean similarity of valid patches $\mathcal{V} = \{i \mid v_i = 1\}$ intersecting the ground-truth crop. Outside Similarity s_c^{out} is the mean similarity over valid patches with no overlap with the crop.

$$s_c^{\text{GT}} = \frac{\sum_{i \in \mathcal{V}} y_{c,i} s_{c,i}}{\sum_{i \in \mathcal{V}}} \quad s_c^{\text{out}} = \frac{\sum_{i \in \mathcal{V}} \mathbf{1}[y_{c,i} = 0] s_{c,i}}{\sum_{i \in \mathcal{V}} \mathbf{1}[y_{c,i} = 0]} \tag{A2}$$

Their difference *Sim Contrast* measures how clearly the representation separates the target region from the remainder of the image:

$$\text{Contrast}_c = s_c^{\text{GT}} - s_c^{\text{out}}. \tag{A3}$$

Recall@k Let $\text{TopK}_c(k) \subseteq \mathcal{V}$ contain the k valid patches with the highest similarities to the crop representation. We define

$$\text{Recall@}\, k_c = \mathbf{1}\left[\max_{i \in \text{TopK}_c(k)} y_{c,i} > 0\right], \qquad k \in \{1, 10\}. \tag{A4}$$

Thus, Recall@k indicates whether at least one ground-truth patch is retrieved among the top-k results.

IoU Let the ground-truth crop have height H_c and width W_c, and let the full-image feature grid have patch strides h_p and w_p. We construct candidate windows of size

$$h_c = \left\lceil \frac{H_c}{h_p} \right\rceil, \qquad w_c = \left\lceil \frac{W_c}{w_p} \right\rceil$$

and slide them over all valid positions of the similarity map. For each candidate window W, its score is the mean similarity of the valid patches it contains:

$$q_c(W) = \frac{\sum_{i \in W \cap \mathcal{V}} s_{c,i}}{\sum_{i \in W \cap \mathcal{V}}}.$$

The highest-scoring window determines the predicted crop location,

$$W_c^* = \arg\max_{W \in \mathcal{W}_c} q_c(W), \tag{A5}$$

where $\mathcal{W}_c$ is the set of candidate windows of size $h_c \times w_c$. After mapping W_c^* to the full-image coordinate system, we obtain a predicted box with the ground-truth crop size and compute IoU. All metrics are averaged over the valid crops in $\mathcal{C}$.

C.2 Other Evaluation Settings for Dental-Arch Awareness

For comparisons with generic SSL methods, we initialize each baseline from its publicly released checkpoint and fine-tune it on the same YoungDent and AIDC training data used for our method. All qualitative and quantitative comparisons use the same held-out test split. Input images are processed at a common resolution of 448×448 for evaluation. For ablation studies, we train model variants with different components and settings on YoungDent only. For in-domain tooth segmentation, all generic SSL baselines, our method, and end-to-end segmentation models use YoungDent only. Because ablation and tooth-segmentation experiments are trained solely on YoungDent, we use Swin-T as the backbone for our method in these settings.

Feature probing settings The k-NN classifier uses cosine similarity with temperature-weighted soft voting ($k = 5$, temperature 0.07). The logistic probe is a class-balanced multinomial logistic regression on L2-normalized features. Its regularization strength is chosen over a subset of the training set, after which the probe is refitted on the full training set and evaluated on the test partition.

Cross-subject FDI Relation Settings We enumerate every pair of test-partition single-tooth crops from different subjects, compute the cosine similarity of their L2-normalized features, and pair it with predefined FDI relation score $R(a, b)$ based on the tooth FDI codes contained within the crops. We exclude pairs that combine single-tooth crops from the frontal, right, or left view with those from the occlusal view, as the occlusal view differs substantially in appearance from the other views. The mean similarity is reported for each FDI relation level. The Spearman correlation is computed over the observed cells of the aggregated 32×32 FDI mean-similarity matrix against the FDI relation matrix.

Representation visualization For PCA visualization, one shared PCA reference pool is built for each model from the dense features output by the training set, balanced by view and drawn from full-mouth images and clean crops. For a sampled full-mouth image and related crops, we render the 3 main PCA components of its patch representations as the feature map. For UMAP visualization, the global representations of sextant and tooth crops are embedded with UMAP, grouped by different views.

C.3 Tooth Segmentation Settings

Datasets We split the in-domain dataset YoungDent at a train-test ratio of 80/20. Given the predicted tooth position by SegmentAnyTooth (Nguyen et al., 2025), each tooth box is used to prompt SAM-HQ (Ke et al., 2023) to generate a pixel-wise tooth mask. For the out-of-distribution (OOD) set, we use the Odontify dataset (Aranguiz & Tobar, 2026), which is never used for either pretraining or training of the segmentation heads in this task.

Segmentation heads & end-to-end models We use segmentation heads and end-to-end models to compare the performance of tooth segmentation with and without our pretraining stage. For the segmentation head, a trainable linear head and a decoder head are designed to be built on the frozen encoders pretrained by our method and other baseline SSL methods. The lightweight linear head ($\sim$ 0.03M parameters) is just a linear projection of features to predicted classes. For the decoder head ($\sim$ 5.9-6.3M parameters), we employ a shared segmentation decoder across all pretrained encoders to ensure a consistent comparison. The encoder features are first projected and resampled into a unified multi-scale feature pyramid, followed by contextual feature aggregation and top-down multi-scale fusion. The fused high-resolution features are then processed by a lightweight prediction head to produce the segmentation logits, which are upsampled to the input resolution. The decoder architecture and training configuration are kept identical across all methods.

For better reference, we also fine-tune end-to-end segmentation models to compare with our segmentation results that use pretraining. We use two different segmentation models, including DeepLabV3 (Chen et al., 2017) with a ResNet-50 backbone initialized from MS-COCO (Lin et al., 2014), and Swin-UPerNet (Liu et al., 2021) pretrained on ImageNet (Deng et al., 2009) directly for segmentation rather than pretraining. Notably, these end-to-end segmentation models have significantly larger trainable parameters compared with the previous setting of frozen pretrained encoders with segmentation heads. Specifically 39.6M for DeepLabV3 and 59.9M for Swin-UPerNet.

Training Both segmentation heads and end-to-end models use the same objective.To train segmentation heads, we optimize with AdamW (lr 3×10^{-4}, weight decay 0.05) and a polynomial scheduler with 500 warm-up steps, at an input resolution of 448×448, batch size 8 per GPU. We set $8,000$ iterations for training with 10% of data and $12,000$ iterations for training with complete data. To train segmentation models, the input resolution is 512×512 to align with the settings in these end-to-end models. We set 4k (10% data) and 40k (100% data) iterations.

Metrics The 32 FDI teeth are grouped into 8 tooth-category classes T_1–T_8 by merging the upper, lower, left and right counterparts of the same tooth category (*e.g.* T_1 = 11/21/31/41, T_2 = 12/22/32/42, …, T_8 = 18/28/38/48). mIoU, mDice and mAcc are the averages over these 8 groups.

D Additional Experiment Results

D.1 More Results on Dental-Arch Awareness Comparison

Comparison of different representation spaces In Tab. 2, localization and probing use pre-projection backbone features to ensure a fair comparison with generic SSL methods. Tab. A2 further compares backbone and projection-head outputs for our Swin and ViT models. For both backbones, alignment-related projection heads primarily improve dense localization by suppressing responses outside the target region. For Swin, similarity contrast, Patch R@1, and IoU increase from 0.268 to 0.712, 0.779 to 0.915, and 0.444 to 0.552, respectively. ViT shows the same pattern. Notably, the ViT projection improves localization despite a lower GT similarity ($0.802 \rightarrow 0.760$), because it substantially reduces similarity to non-overlapping regions ($0.599 \rightarrow 0.211$). Thus, localization depends on target–background separation rather than target similarity alone.

In contrast, pre-projection features consistently yield stronger tooth and sextant probing, with the gap particularly noticeable for ViT. Also, FDI relation head significantly expands the FDI relation range when using Swin Transformer as backbone ($0.392 \rightarrow 0.706$). This division of roles is intentional: routing local-to-global alignment and FDI-relation objectives through separate projection heads lets each specialize in its target without allowing either objective to dominate the shared backbone. The backbone therefore retains discriminative, transferable tooth semantics.

Localization across structural crop types Tab. A3 compares localization across single-tooth and sextant crops. Sextant crops are easier to localize for every method, as their larger contiguous regions are less affected by small spatial shifts. However, both variants of our method show a smaller tooth-to-sextant gap than all baselines. In particular, Ours (Swin) increases from 0.465 to 0.772 in IoU ($1.66\times$), compared with at least $4.29\times$ for the baselines, indicating greater robustness to crop scale.

Table A2: Effect of global and patch projection heads. Swin and ViT features are evaluated with and without projection for localization, feature probing, and cross-subject FDI tooth relations. *w/ proj* denotes the projection head output, and *w/o proj* is the pre-projection backbone feature.

	Swin		ViT	
	w/ proj	w/o proj	w/ proj	w/o proj
Localization				
GT Sim. ↑	0.733	0.642	0.760	0.802
Out. Sim. ↓	0.021	0.374	0.211	0.599
Sim. Con. ↑	0.712	0.268	0.549	0.203
Patch R@1/5/10 ↑	0.915/0.990/0.997	0.779/0.902/0.935	0.603/0.812/0.898	0.550/0.710/0.787
IoU ↑	0.552	0.444	0.338	0.280
Probing (KNN k=5 / Logistic, macro-F1)				
Tooth FDI ↑	0.745/0.758	0.754/0.808	0.205/0.382	0.368/0.624
Tooth class ↑	0.822/0.817	0.832/0.851	0.328/0.500	0.475/0.663
Sextant ↑	0.953/0.934	0.954/0.956	0.821/0.898	0.897/0.935
Cross-subject FDI tooth relation (mean ± std)				
R=0.05 (weak relation)	$0.136_{\pm 0.256}$	$0.477_{\pm 0.138}$	$0.512_{\pm 0.364}$	$0.662_{\pm 0.236}$
R=0.25 (same sextant)	$0.353_{\pm 0.231}$	$0.563_{\pm 0.152}$	$0.578_{\pm 0.342}$	$0.713_{\pm 0.219}$
R=0.45 (same category)	$0.333_{\pm 0.269}$	$0.609_{\pm 0.149}$	$0.620_{\pm 0.310}$	$0.738_{\pm 0.201}$
R=0.60 (adjacent)	$0.591_{\pm 0.250}$	$0.703_{\pm 0.158}$	$0.662_{\pm 0.307}$	$0.766_{\pm 0.202}$
R=0.70 (symmetric)	$0.800_{\pm 0.209}$	$0.850_{\pm 0.132}$	$0.659_{\pm 0.294}$	$0.767_{\pm 0.190}$
R=1.00 (same FDI)	$0.841_{\pm 0.210}$	$0.868_{\pm 0.133}$	$0.689_{\pm 0.297}$	$0.785_{\pm 0.196}$
Relation range ↑	0.706	0.392	0.177	0.123
Spearman ↑	0.669	0.674	0.297	0.373

Table A3: Localization by crop type. Single-tooth and sextant crops are evaluated using GT/Outside similarity, similarity contrast, Patch Recall@k, and box IoU.

Model	**Crop**	**GT Sim. ↑**	**Out. Sim. ↓**	**Sim. Con. ↑**	**Patch R@1/5/10 ↑**	**IoU ↑**
Masked Autoencoding						
Swin-MAE	Tooth	0.916	0.899	0.018	0.052/0.166/0.255	0.020
	Sextant	0.928	0.918	0.010	0.129/0.335/0.493	0.113
Contrastive Learning						
DINOv2	Tooth	0.529	0.374	0.155	0.206/0.527/0.682	0.106
	Sextant	0.656	0.440	0.216	0.699/0.944/0.980	0.455
SimCLRv2	Tooth	0.192	0.131	0.061	0.167/0.419/0.592	0.053
	Sextant	0.243	0.147	0.096	0.602/0.844/0.948	0.394
Masked Alignment Pretraining						
Ours (Swin)	Tooth	0.649	0.301	**0.347**	**0.815/0.968/0.984**	**0.465**
	Sextant	0.696	0.383	**0.314**	**0.995/0.997/0.997**	**0.772**
Ours (ViT)	Tooth	0.781	0.623	0.157	0.314/0.515/0.624	0.181
	Sextant	0.893	0.657	0.237	0.947/0.980/0.987	0.699

The advantage is most pronounced for the harder single-tooth queries. Ours (Swin) achieves 0.815 Patch R@1 and 0.465 IoU, exceeding the strongest baseline, DINOv2, by nearly $4.0\times$ and $4.4\times$, respectively. It also obtains the highest similarity contrast for both tooth (0.347) and sextant (0.314) crops. In contrast, Swin-MAE assigns similarly high similarity to target and non-overlapping regions, resulting in minimal contrast (0.018/0.010) and poor localization. These results show that effective localization requires separating the target from irrelevant patches, rather than merely maximizing target similarity.

Feature map visualization In addition to Fig. 5, Fig. A2 extends the visualization to four-tooth group crops that are not used during pretraining. Each case compares a crop with its source full-mouth image. Our methods localize the target regions more clearly, with Ours (ViT) showing particularly concentrated similarity contours, whereas generic SSL baselines exhibit more diffuse or ambiguous

responses. The full-mouth maps also distinguish tooth categories, especially for Ours (ViT), rather than only separating teeth from surrounding tissues like DINOv2. Moreover, crop and source regions retain similar color families while resolving finer details. Swin-MAE produces spatially compressed maps with diverse but messy semantics and weaker tooth-level semantic separation.

Visualization of cross-subject FDI relation Fig. 6(a) reports the cross-subject similarity between pairs of single-tooth crops, computed with the FDI relation head. To inspect how much of this structure the encoder itself already provides, we supplement the same matrix with the similarity obtained from the unprojected backbone features in Fig. A3(a), and likewise report the cross-subject similarity of sextant and tooth category with and without the FDI relation head in Fig. A3(b–c). The projected features form a clearly structured pattern, assigning higher similarity to teeth of the same category and separating the different FDI relation levels, whereas the unprojected backbone features show no comparably distinct pattern. This contrast reflects the importance of the task-specific projection head. The underlying anatomical structure, however, is already present in the backbone. Teeth of the same category remain more similar to one another than to the other teeth, and symmetric sextants remain more similar to each other than to non-symmetric ones.

UMAP visualization of single-tooth crop feature distributions In addition to Fig. 7 and Fig. 6(b), we present the UMAP feature distributions of single-tooth crops across different SSL methods. As shown in Fig. A4, we visualize the features extracted from the upper occlusal and frontal views. Since tooth structure and appearance vary significantly across viewpoints, especially in occlusal views, we partition the distributions by view. Given their distinct anatomical representations, we select these two views as representative cases. For the upper occlusal view, we include tooth categories 1 through 8. For the frontal view, we restrict the visualization to tooth categories 1 through 3, as premolars and molars often have limited or no visibility. Notably, both our model and SimCLRv2 demonstrate clear tooth-level feature separation, whereas the features of different tooth categories remain largely intertwined in DINOv2 and Swin-MAE.

D.2 More Results on Tooth Segmentation

Tab. A4 and Tab. A5 give the per-class tooth segmentation IoU behind Tab. 5 for in-domain settings. The former compares frozen pretrained encoders with trainable segmentation heads, whereas the latter juxtaposes Ours (Swin) with a frozen encoder and decoder against end-to-end fine-tuned segmentation models. In both tables the 32 FDI classes are merged into 8 tooth-category groups T_1–T_8 by aggregating the upper, lower, left and right counterparts of each tooth category, *i.e.* T_1 = 11/21/31/41 and T_8 = 18/28/38/48 denote the central incisors and the third molars (see Sec. C.3). mIoU, mDice and mAcc are the averages over these 8 groups with the background excluded.

Effect of the segmentation head and of the data scales For every pretrained encoder, replacing the linear head by the decoder yields a large gain. Swin-MAE improves from 0.094 to 0.535 mIoU under 10% labels and from 0.119 to 0.614 under 100% labels. SimCLRv2 from 0.220 to 0.396 and from 0.256 to 0.448. DINOv2 from 0.257 to 0.510 and from 0.291 to 0.574. The improvement is smallest for our Swin model ($0.392 \rightarrow 0.634$ and $0.411 \rightarrow 0.671$), which is expected since only our frozen features already support a strong linear probe, being the sole method whose linear head stays above 0.39 mIoU under training data scales. In contrast, increasing the training data from 10% to 100% changes the result far less for these frozen encoders, adding at most 0.079 mIoU (our Swin: $+0.019$ with the linear head and $+0.037$ with the decoder), whereas the end-to-end models gain 0.099 (DeepLabV3) and 0.161 (Swin-UPerNet). The representation quality obtained in the pretraining stage is therefore the dominant factor once a decoder is attached, and the training data of the downstream segmentation task is comparatively secondary.

Fig. A5 and A6 show the visualization of tooth segmentation performance in different segmentation heads and data scales using the in-domain YoungDent dataset. Fig. A7 further shows visualization examples from the out-of-domain Odontify dataset.

Comparison with end-to-end segmentation models Our Swin model with the decoder attains 0.634 mIoU (0.754 mDice, 0.766 mAcc) using only 10% of the labeled data, already above DeepLabV3 (0.621) and Swin-UPerNet (0.510) trained on the same data size, while consuming no dense annotation in the pretraining stage. With the full dataset it reaches 0.671 mIoU (0.789 mDice,

Model	Decoder	100%	T1	T2	T3	T4	T5	T6	T7	T8	mIoU	mDice	mAcc
Swin-MAE	✗	✗	0.164	0.103	0.149	0.091	0.097	0.117	0.028	0.000	0.094	0.167	0.180
	✗	✔	0.181	0.129	0.166	0.132	0.120	0.157	0.068	0.000	0.119	0.208	0.235
	✔	✗	0.760	0.717	0.691	0.569	0.469	0.484	0.400	0.190	0.535	0.678	0.697
	✔	✔	0.840	0.794	0.762	0.645	0.554	0.567	0.477	0.275	0.614	0.745	0.795
SimCLRv2	✗	✗	0.359	0.304	0.253	0.235	0.221	0.214	0.138	0.039	0.220	0.351	0.408
	✗	✔	0.409	0.348	0.288	0.278	0.250	0.272	0.173	0.034	0.256	0.396	0.461
	✔	✗	0.617	0.533	0.499	0.448	0.368	0.381	0.273	0.051	0.396	0.545	0.583
	✔	✔	0.659	0.572	0.539	0.485	0.403	0.428	0.324	0.172	0.448	0.604	0.698
DINOv2	✗	✗	0.370	0.381	0.310	0.259	0.242	0.293	0.191	0.008	0.257	0.394	0.482
	✗	✔	0.410	0.425	0.358	0.282	0.277	0.320	0.242	0.017	0.291	0.436	0.537
	✔	✗	0.763	0.698	0.664	0.557	0.448	0.472	0.376	0.104	0.510	0.649	0.674
	✔	✔	0.805	0.745	0.712	0.608	0.511	0.543	0.437	0.233	0.574	0.712	0.773
Ours (Swin)	✗	✗	**0.538**	**0.519**	**0.481**	**0.407**	**0.396**	**0.419**	**0.305**	**0.070**	**0.392**	**0.546**	**0.648**
	✗	✔	**0.557**	**0.532**	**0.495**	**0.435**	**0.409**	**0.432**	**0.326**	**0.104**	**0.411**	**0.567**	**0.667**
	✔	✗	**0.861**	**0.825**	**0.807**	**0.707**	**0.589**	**0.585**	**0.491**	**0.209**	**0.634**	**0.754**	**0.766**
	✔	✔	**0.874**	**0.837**	**0.825**	**0.720**	**0.623**	**0.633**	**0.530**	**0.329**	**0.671**	**0.789**	**0.830**
Ours (ViT)	✗	✗	0.421	0.365	0.364	0.320	0.165	0.300	0.234	0.000	0.271	0.409	0.464
	✗	✔	0.475	0.421	0.398	0.343	0.214	0.336	0.259	0.006	0.306	0.449	0.519
	✔	✗	0.769	0.697	0.669	0.568	0.444	0.461	0.380	0.181	0.521	0.665	0.691
	✔	✔	0.816	0.755	0.719	0.619	0.508	0.527	0.438	0.246	0.578	0.716	0.777

Table A4: In-domain tooth segmentation with frozen pretrained encoders. Per-position IoU and mIoU/mDice/mAcc are reported for linear/decoder heads and 10%/100% labeled data; ✔ denotes the decoder head or 100% data. Bold indicates the best result for each head and data-scale setting. T_1–T_8 merge the upper, lower, left, and right counterparts of each tooth category.

Model	100%	T1	T2	T3	T4	T5	T6	T7	T8	mIoU	mDice	mAcc
DeepLabV3	✗	0.839	0.794	0.778	0.673	0.582	**0.589**	0.466	**0.245**	0.621	0.748	0.747
DeepLabV3	✔	**0.908**	**0.881**	**0.863**	**0.778**	**0.692**	**0.696**	**0.582**	**0.357**	**0.720**	**0.824**	0.820
Swin-UPerNet	✗	0.641	0.626	0.643	0.560	0.496	0.489	0.387	0.238	0.510	0.664	0.656
Swin-UPerNet	✔	0.852	0.821	0.816	0.737	0.631	0.645	0.528	0.337	0.671	0.790	0.780
Ours (Swin) + Decoder	✗	**0.861**	**0.825**	**0.807**	**0.707**	**0.589**	0.585	**0.491**	0.209	**0.634**	**0.754**	**0.766**
Ours (Swin) + Decoder	✔	0.874	0.837	0.825	0.720	0.623	0.633	0.530	0.329	0.671	0.789	**0.830**

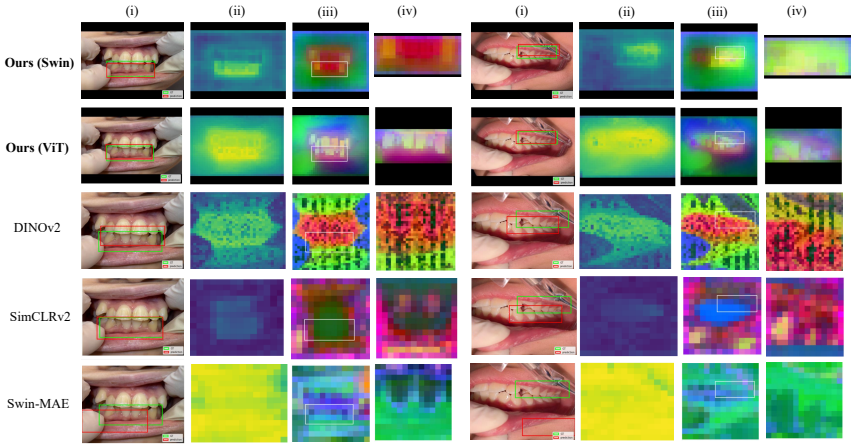
Table A5: In-domain tooth segmentation: frozen Ours versus end-to-end models. Per-position IoU and mIoU/mDice/mAcc are reported for 10% (✗) and 100% (✔) labeled training data. Bold indicates the best result at each data scale.

0.830 mAcc), matching Swin-UPerNet (0.671) and staying 0.049 below DeepLabV3 (0.720). The gap is thus concentrated in the high-label regime, and our frozen-representation pipeline is clearly the more label-efficient option when dense labels are scarce. Fig. A8 shows a consistent pattern.

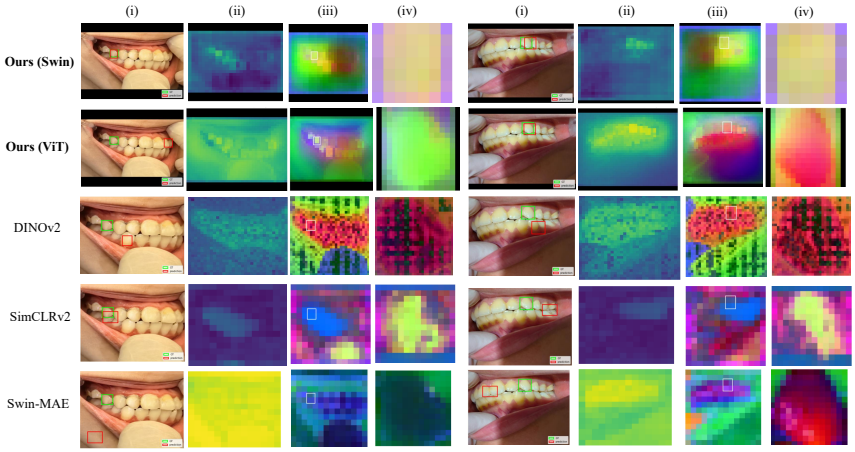
Per-class behavior All methods exhibit an overall decline in IoU from the anterior to the posterior groups, although adjacent categories may show small fluctuations. Taking our Swin model with the decoder and 100% labels, T_1–T_3 remain above 0.82 (0.874/0.837/0.825), T_4–T_6 drop to 0.720/0.623/0.633 and T_7/T_8 to 0.530/0.329. The end-to-end models follow the same trend, from 0.908 to 0.357 for DeepLabV3 and from 0.852 to 0.337 for Swin-UPerNet under 100% labels. We attribute this to the smaller size of the posterior teeth, their frequent occlusion and truncation at the image boundary in buccal and occlusal views, and their mutual visual similarity, notably between the second and third molars. Even in this hardest regime, Ours (Swin) remains the strongest frozen pretrained encoder: on T_8 with 100% labels, it reaches 0.329, compared with 0.275 (Swin-MAE), 0.233 (DINOv2), 0.172 (SimCLRv2), and 0.246 (Ours (ViT)).

E LIMITATIONS AND BROADER IMPACT

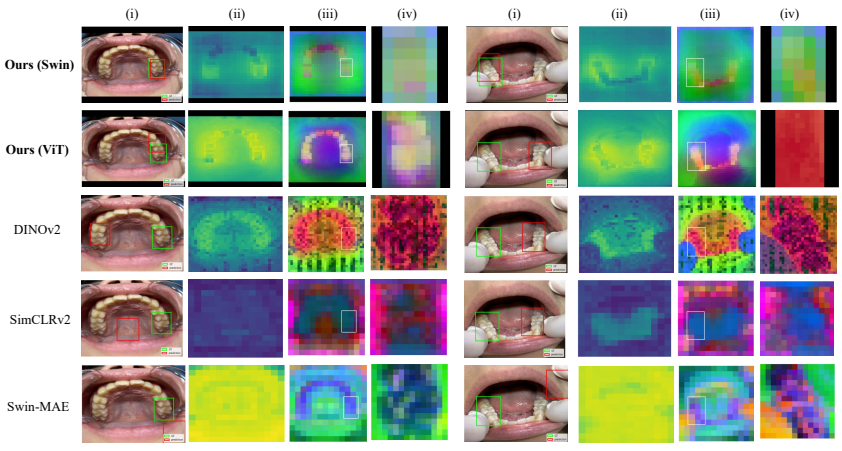
Our structural crops and FDI-aware objectives rely on an upstream tooth-numbering model to provide tooth boxes and FDI codes. Imperfect localization, partial visibility, or coding errors may therefore propagate to the crop construction and relation supervision. This could be reduced by incorporating confidence-aware filtering and human-verified annotations. Our YoungDent dataset provides a relatively general multi-view and multi-device setting for intraoral representation learning, while our experiments still cover a limited number of clinical sources and populations. Broader validation across clinical sites and demographic groups is therefore needed to further assess robustness. The learned representations may support more downstream applications such as oral conditions diagnosis such as caries, gingivitis, and dental plaque. Deployment should also follow applicable privacy and data-governance requirements, since intraoral images may contain sensitive health information, and performance should be audited across demographic groups and acquisition settings to identify potential biases.



(a) Sextant crop.



(b) Single-tooth crop.



(c) Four-tooth group crop, unseen during pretraining.

Figure A2: Qualitative comparison of target crop localization and local-to-global feature consistency. Within each panel, (i)–(iv) show the predicted/GT bounding boxes of the target crop, crop-to-image similarity heatmap, full-mouth feature map, and target-crop feature map, respectively.

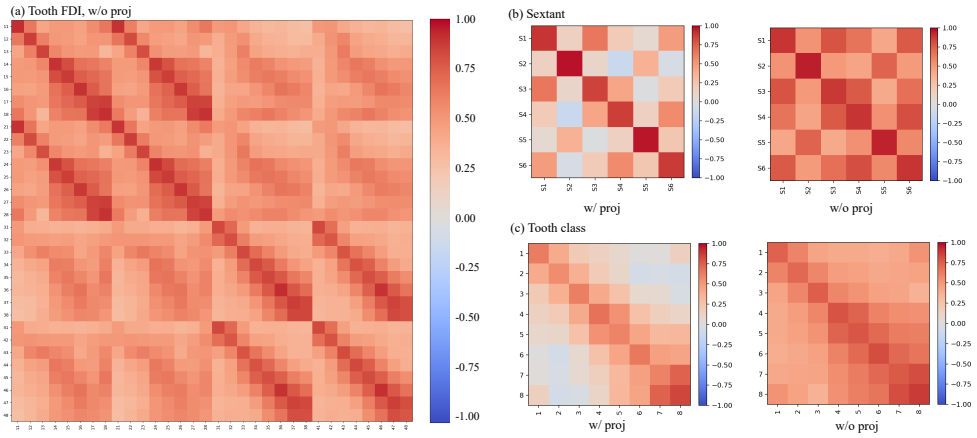


Figure A3: Cross-subject similarity matrices with and without FDI relation projection. Each cell is the mean cosine similarity for the corresponding FDI, sextant, or tooth-category pair. (a) Single-tooth backbone features indexed by FDI code. (b,c) Sextant and tooth-category features after (left) and before (right) FDI relation projection.

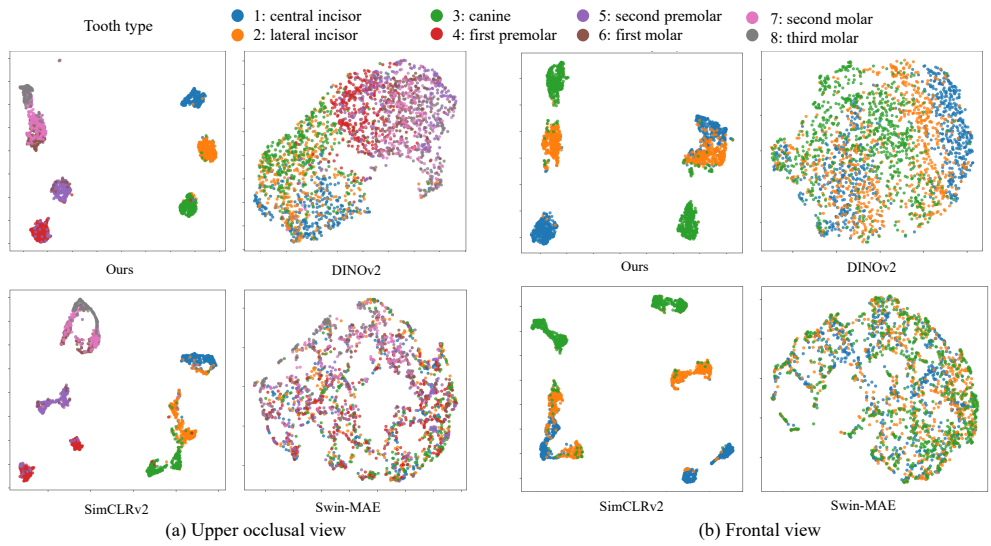


Figure A4: UMAP of single-tooth crop representations. Embeddings from different SSL encoders are visualized for (a) upper-occlusal and (b) frontal views; colors denote tooth category.

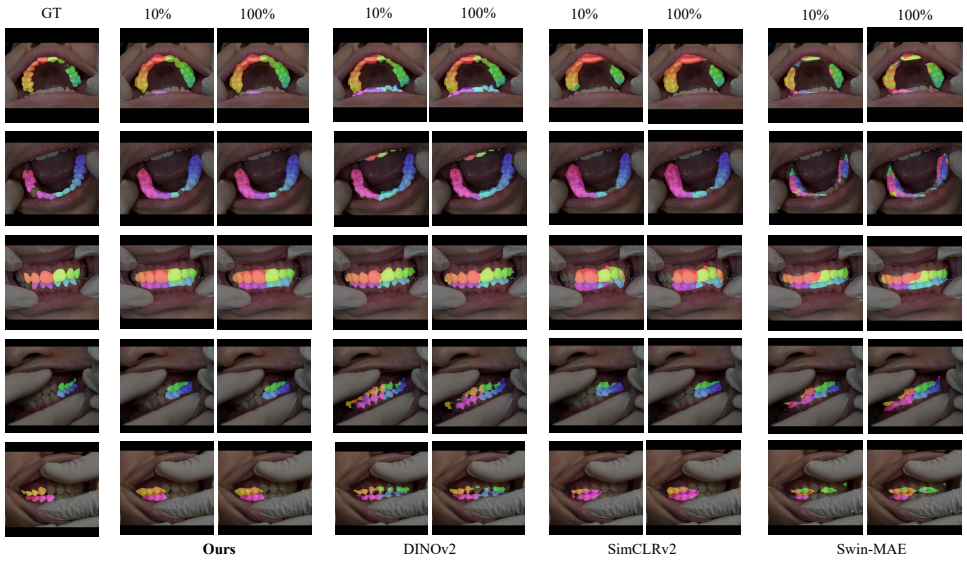


Figure A5: In-domain tooth segmentation with frozen encoders and linear heads. Predictions use 10% and 100% of labeled YoungDent training data.

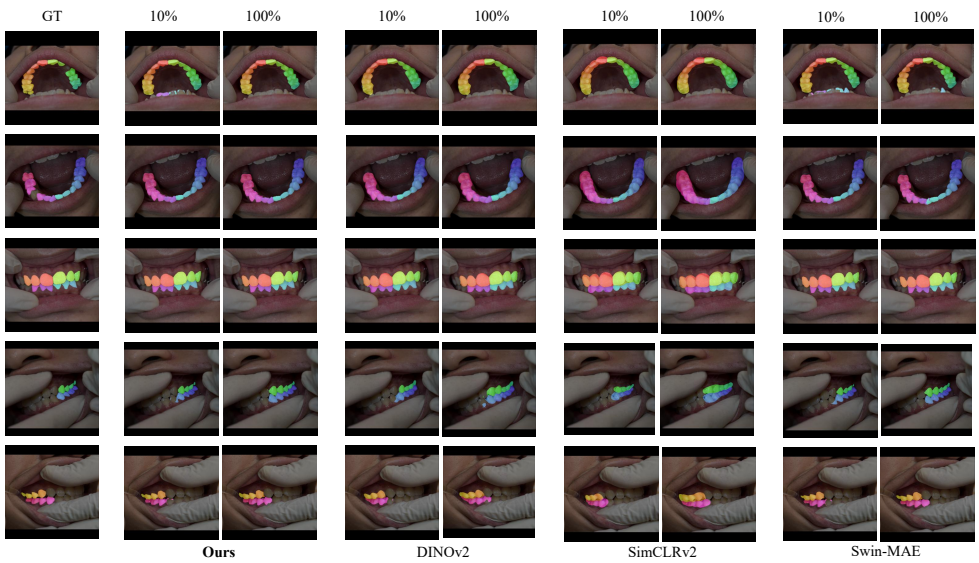


Figure A6: In-domain tooth segmentation with frozen encoders and decoder heads. Predictions use 10% and 100% of labeled YoungDent training data.

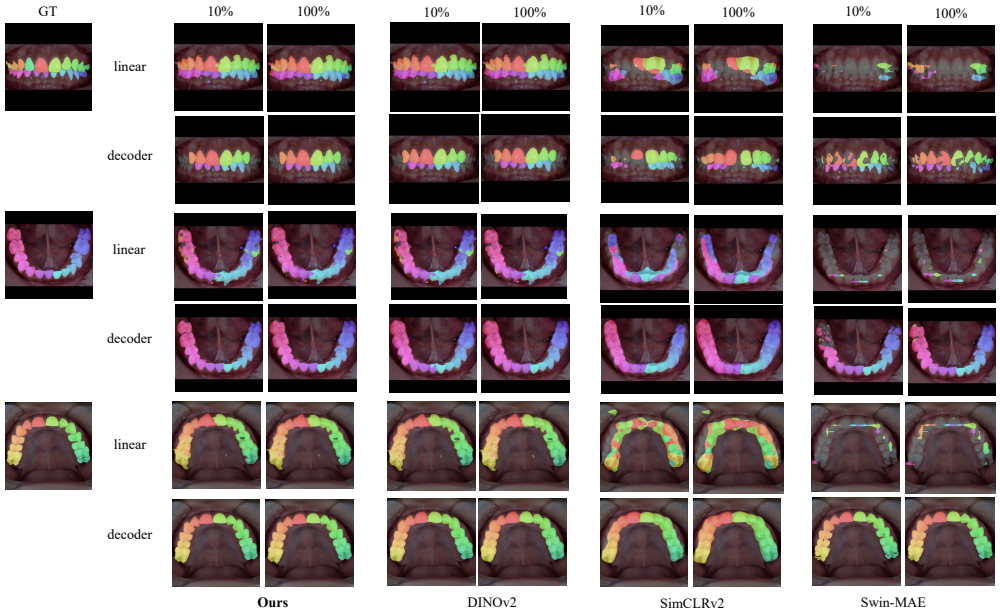


Figure A7: Out-of-domain tooth segmentation on Odontify (Aranguiz & Tobar, 2026). Frozen SSL encoders are evaluated with linear and decoder heads trained using 10% or 100% of labeled YoungDent data.

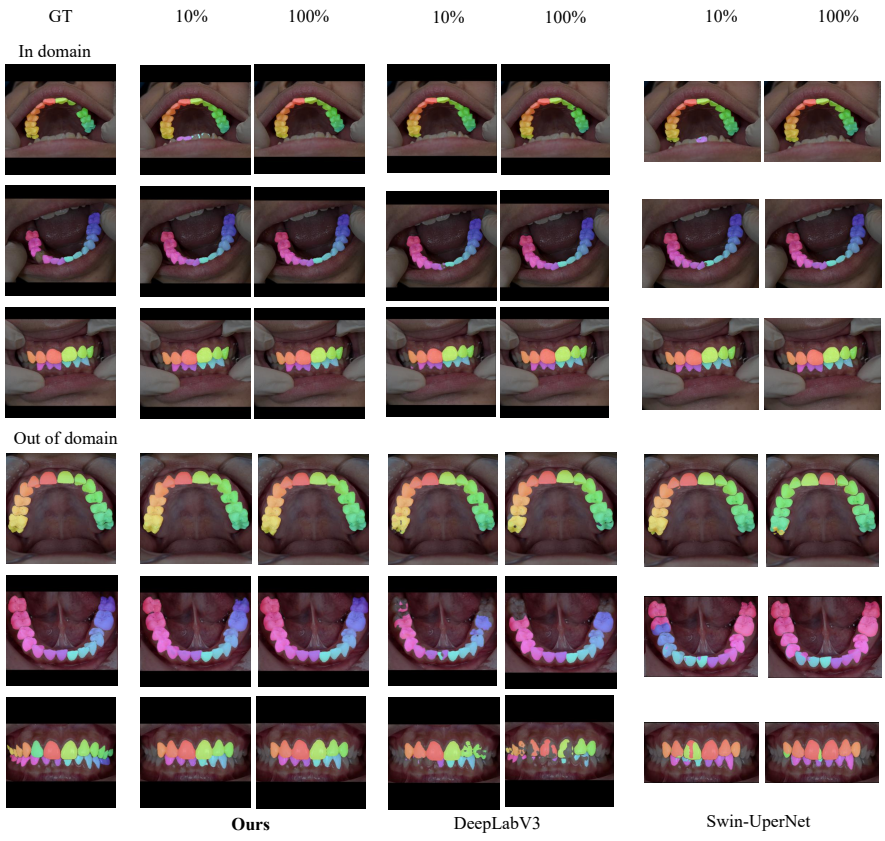


Figure A8: Comparison with end-to-end segmentation models. Our frozen pretrained encoder with a trainable decoder is compared with DeepLabV3 and Swin-UPerNet on YoungDent and Odontify.