Article

Not peer-reviewed version

Oral Disease Recognition via Dynamic Compositional Prototype Learning

Wenyuan Wang , Ruisi Yang , Jutao Xiao , Shuwei Huo * , Zhengze Chen

Posted Date: 1 April 2025

doi: 10.20944/preprints202504.0003.v1

Keywords: oral disease recognition; prototype learning; dynamic compositional prototype

 doi:10.20944/preprints202504.0003.v1

Article

Oral Disease Recognition via Dynamic Compositional Prototype Learning

Wenyuan Wang [1,†], Ruisi Yang [2,†], Jutao Xiao [3], Shuwei Huo [3,*] and Zhengze Chen [4]

[1] School of Management and Engineering, Capital University of Economics and Business, Beijing 100070, China
[2] School of Stomatology, Hebei Medical University, Shijiazhuang 050017, China
[3] School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
[4] College of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin 300130, China
* Correspondence: huosw@cse.neu.edu.cn
† These authors contributed equally to this work. Author order was determined alphabetically by surname.

Abstract: Oral disease recognition aims to identify specific diseases from individual oral images. With the progress of deep learning, pioneering computational methods tailored for this task have started to demonstrate their potential. Conventional deep learning approaches typically treat oral disease recognition as a simplistic image-to-label mapping, thereby neglecting the critical need to explicitly model disease-specific visual patterns. This oversimplification compromises their ability to generalize effectively to unseen data, particularly when training data is limited. To address this issue, we analyzed the mechanisms of the diagnostic process, which involves first identifying pathological features in images and then determining the corresponding oral disease based on these features. *Current methods, however, overlook the step of extracting pathological features, resulting in suboptimal model performance.* **To overcome this limitation, we propose a novel framework termed Dynamic Compositional Prototype Learning (DyCoP).** The DyCoP framework leverages componential prototypes and category prototypes to represent localized pathological features and oral diseases, respectively. It employs a dynamic composition mechanism to establish relationships between multiple pathological features and specific oral diseases. Furthermore, we introduce a gradient suppression strategy to fully utilize general knowledge embedded in pretrained models. These approaches enhance the model's capacity to capture effective features and learn accurate diagnostic logic. Experiments conducted on the Dental Condition Dataset demonstrate the superiority of our method, achieving an accuracy of 93.3% and a macro-F_1 score of 91.9%, outperforming state-of-the-art approaches by significant margins. This framework effectively addresses both precision and generalizability bottlenecks in automated oral disease diagnosis.

Keywords: oral disease recognition; prototype learning; dynamic compositional prototype

1. Introduction

Oral disease recognition focuses on detecting specific oral disease from single oral images, helping identify issues like calculus or gingivitis. With the development of deep learning technology, researchers are now developing smarter computer programs that can analyze these images effectively, showing promising results in early testing [1–3]. These AI-driven tools may hold the potential to improve access to affordable dental screenings, especially in remote or underserved communities where specialist care is scarce.

Existing methods usually consider oral disease recognition as a simple classification task [4–6], using deep neural networks [7–13] to map images directly to predefined labels. However, this approach fails to capture the intrinsic characteristics of oral diseases, as it ignores explicit modeling of pathological features. This limitation leads to poor generalization, particularly with limited or low-quality data. To address this, we analyzed the diagnostic process of experienced clinicians, who follow a structured two-stage approach (Figure 1): first identifying localized pathological features (e.g.,

 doi:10.20944/preprints202504.0003.v1

calculus deposits, gingival bleeding points, or root caries), then cross-referencing these observations with established diagnostic criteria through multidimensional comparisons. In contrast, current deep learning methods lack mechanisms to incorporate such critical pathological reasoning, relying instead on superficial image-to-label mappings that risk learning biased correlations.

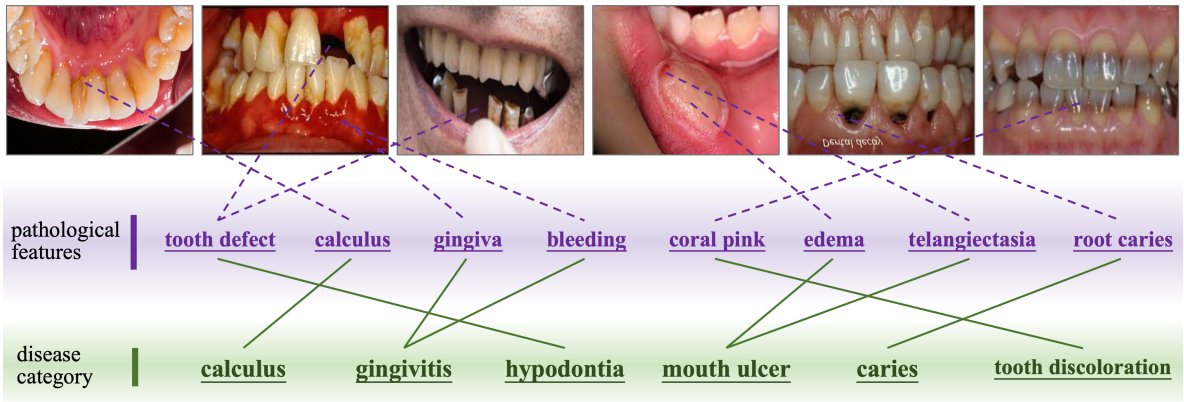

Figure 1. An example of the diagnostic process for oral diseases: initially pinpointing pathological features—such as calculus deposits, gingival bleeding points, or root caries—and subsequently cross-referencing these observations with established disease indicators through multi-dimensional comparisons.

To tackle these issues, we propose Dynamic Compositional Prototype Learning (DyCoP), a framework designed to emulate clinical diagnostic logic through three innovations:

- **Compositional Disease Representation**. DyCoP decomposes complex diseases into reusable componential prototypes representing fundamental visual patterns shared across oral pathologies (e.g., "calculus deposits" or "enamel demineralization"). This decomposition enables efficient knowledge transfer, reducing reliance on large datasets.
- **Dynamic Prototype Assembly**. A sparse activation mechanism adaptively assembles disease-specific signatures from relevant prototypes based on input image characteristics. This flexibility accommodates intra-class variations across disease stages and presentations.
- **Strategic Preservation of Pre-trained Knowledge**. By performing gradient suppression to lower-level convolutional layers during training, DyCoP retains generalized visual feature extraction capabilities from pre-trained models while fine-tuning higher layers for oral disease specifics. This balances domain adaptation with data efficiency.

Experiments on the Dental Condition Dataset—a curated collection of annotated oral images—demonstrate DyCoP's superiority over state-of-the-art models, achieving 93.3% accuracy and 91.9% macro-F1 score across six common oral diseases (calculus, gingivitis, hypodontia, mouth ulcer, caries, and tooth discoloration). The framework's emphasis on explicit pathological feature modeling also enhances interpretability, aligning activated prototypes with clinically meaningful regions. Beyond oral diseases, DyCoP offers a adaptable framework for other medical domains facing similar data and complexity challenges.

2. Related Work

Our main work is to propose a novel prototype learning method for oral disease recognition. In this section, we review related works, including deep learning-based oral disease recognition and prototype learning methods for image analysis.

2.1. Oral Disease Recognition

Recent advances in deep learning have revolutionized oral disease diagnostics by addressing limitations of traditional methods. Hybrid architectures like InceptionResNetV2 [4,6] and optimization techniques such as Common Vector Approach (CVA)-pool weighting [5] enhance feature extraction and classification accuracy for caries, gingivitis, and oral lesions. AI-driven biomarker discovery [14]

 doi:10.20944/preprints202504.0003.v1

enables non-invasive salivary analysis, while specialized models demonstrate precision in detecting recurrent aphthous ulcers [15] and oral cancers [16,17]. Systematic reviews [18,19] validate AI's clinical potential, particularly in radiographic interpretation and lesion classification. Innovations in model interpretability [17], multi-modal data integration [20], and ensemble strategies [21,22] highlight the shift toward transparent, clinically actionable systems. These developments collectively establish AI as a transformative tool across diagnostic workflows, from early detection to treatment planning, while frameworks like ACES [23] (Application of the 2018 periodontal status Classification to Epidemiological Survey data) provide standardized evaluation protocols for periodontal conditions. *Current deep learning methods often simplify diagnosis to a basic image-to-label mapping, overlooking the need to model disease-specific visual patterns explicitly. This limits their generalization to new data, especially with scarce training samples. We introduce the DyCoP framework, utilizing componential and category prototypes to capture localized pathological features and oral diseases, addressing these shortcomings.*

2.2. Prototype Learning for Image Analysis

Prototype learning represents a classical method in pattern recognition where prototypes serve as representatives for sets of examples. These prototypes can be obtained by designing updating rules [24] or minimizing loss functions [25]. In semi-supervised learning tasks [26–28], prototypes represent the data manifold for each category, facilitating supervised information propagation and noisy label correction. For image classification, prototype learning approaches [25,29] address recognition challenges by assigning multiple prototypes to different classes, enhancing classification robustness through prototype-based decision functions and distance computations. The application of prototype learning has expanded to weakly supervised or few-shot semantic segmentation [30–34], few-shot object detection [35,36], and person re-identification [37–40]. The prototype mixture model (PMM) [32] enforces prototype-based semantic representation from limited support images by correlating diverse image regions with multiple prototypes. To improve weakly supervised semantic segmentation (WSSS), various prototype-guided solutions have emerged. An unsupervised principal prototypical features discovering strategy was developed [30] for initial object localization, though these prototypes cannot be learned end-to-end. To generate more precise segmentation pseudo masks, cross-view feature semantic consistency regularization was implemented [33] using pixel-to-prototype contrast while promoting intra-class compactness and inter-class dispersion in the feature space. *Current prototype learning methods focus solely on establishing prototypes at the category level, which limits their representational capacity. We propose a hierarchical prototype approach, constructing prototypes at both the component level and the category level to enhance the method's representational power.*

3. Methodology

In this section, we describe the proposed DyCoP framework. As shown in Figure 2, DyCoP recognizes an oral condition image through four steps: 1) feature extraction, which receives an image and represents it as a feature vector $x \in \mathbb{R}^d$; 2) componential prototype matching, which matches the feature x with a series of componential prototypes to identify which semantic components are present, resulting in a set of weights; 3) compositional representation generation, which creates a compositional image representation based on these weights; 4) finally, category prototype matching (classification), which generates classification results by comparing the similarity between the compositional representation and category prototypes.

 doi:10.20944/preprints202504.0003.v1

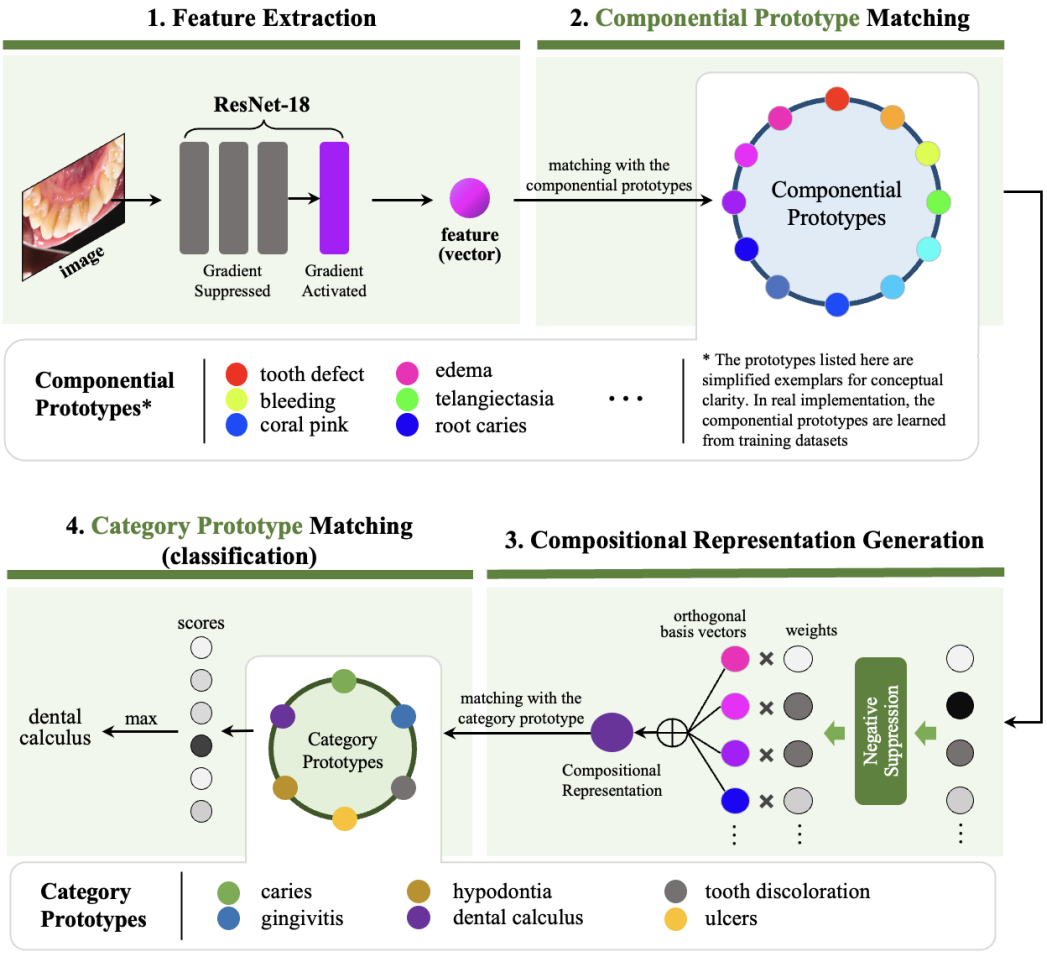

Figure 2. Overview of our proposed **Dy**namic **Co**mpositional **P**rototype (DyCoP) framework. The DyCoP recognizes an oral condition image through four steps: 1) feature extraction, which receives an image and represents it as a feature vector; 2) componential prototype matching, which matches the feature with a series of componential prototypes to identify which semantic components are present, resulting in a set of weights; 3) compositional representation generation, which creates a compositional image representation based on these weights; 4) finally, category prototype matching (classification), which generates classification results by comparing the similarity between the compositional representation and category prototypes.

3.1. Feature Extraction

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, we employ a parametric mapping function $\phi_\theta : \mathcal{I} \to \mathcal{X}$ implemented as a ResNet-18 architecture to obtain discriminative feature vector $x \in \mathbb{R}^d$, where d denotes the vector dimension. The feature extractor is initialized with weights pre-trained on ImageNet (ILSVRC 2012), inheriting powerful visual pattern recognition capabilities.

To preserve the hierarchical feature representations learned from large-scale datasets while adapting to downstream tasks, we propose a progressive gradient annealing strategy. Specifically, for the first L convolutional blocks containing low-level edge detectors and texture filters, we perform **gradient suppression strategy** during backpropagation:

$$\mathbb{E}_{(\mathbf{I},y)\sim\mathcal{D}_{\text{train}}}\left[\frac{\partial\mathcal{L}_{\text{task}}}{\partial\theta_l}\right] = \begin{cases} 0, & \forall l \in \{1,\ldots,L\} \\ \mathbb{E}\left[\nabla_{\theta_l}\mathcal{L}_{\text{task}}\right], & \text{otherwise} \end{cases} \tag{1}$$

where θ_l parameterizes the l-th convolutional layer, and $\mathcal{L}_{\text{task}}$ denotes the loss function used in the training process of this work.

 doi:10.20944/preprints202504.0003.v1

3.2. Componential Prototype Matching

We construct a parametric prototype set $P \triangleq \{p_k\}_{k=1}^{K} \subset \mathbb{R}^d$ where each basis vector p_k represents a componential prototype (atomic visual concept). The component dictionary is implemented as a trainable parameter matrix $P \in \mathbb{R}^{K \times d}$ with Xavier initialization:

$$p_k^{(0)} \sim \mathcal{N}\left(0, \sqrt{\tfrac{2}{d}}\, I_d\right) \tag{2}$$

where d denotes the dimensions of the feature vector, and I_d is a $d \times d$ identity matrix. The prototype matrix undergoes gradient updates through $P^{(t+1)} \leftarrow P^{(t)} - \eta_t \nabla_P \mathcal{L}_{\text{orth}} + \lambda(P^{(t)} - P^{(t-1)})$ where η_t is the adaptive learning rate and λ controls momentum retention. The orthogonality constraint $\mathcal{L}_{\text{orth}} = \|PP^\top - I_K\|_F^2$ ensures the prototype vectors are linearly independent.. After sufficient iterations, we can establish a set of orthogonal componential prototypes, which remain trainable during subsequent training processes.

Given an oral disease image feature vector $\mathbf{x} \in \mathbb{R}^d$, we quantify its semantic alignment with componential prototypes through an adaptive similarity metric. The matching coefficient $\alpha_k \in \mathbb{R}_{\geq 0}$ for the k-th componential prototype $\mathbf{p}_k \in \mathbb{R}^d$ is computed via:

$$\alpha_k = \mathcal{T}_{\text{NS}}\left(\mathbf{x}^\top \mathbf{p}_k\right) \tag{3}$$

where $\mathcal{T}_{\text{NS}} : \mathbb{R} \to \mathbb{R}_{\geq 0}$ denotes the non-linear negative suppression function defined as $\mathcal{T}_{\text{NS}}(z) = \max(0, z)$. This piecewise-linear transformation implements selective feature emphasis by suppressing non-positive values while preserving positive ones. The resulting coefficients $\{\alpha_k\}_{k=1}^{K}$ form a sparse attention distribution over prototypes, encoding discriminative visual patterns of the oral diseases.

3.3. Dynamic Compositional Representation Generation

Let $\mathcal{E} = \{\mathbf{e}_k\}_{k=1}^{K}$ denote the canonical orthonormal basis for $\mathbb{R}^K$, uniquely characterized by the orthonormality condition:

$$\langle \mathbf{e}_k, \mathbf{e}_j \rangle = \delta_{kj}, \quad \forall 1 \leq k, j \leq K \tag{4}$$

where δ_{kj} represents the Kronecker delta whose value is 1 when $k = j$ or 0 otherwise, and $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product. The basis satisfies the fundamental duality relationship: $\mathbf{e}_k^\top \mathbf{x} = x_k, \forall \mathbf{x} \in \mathbb{R}^K$.

The compositional representation is generated through sparse linear combination of basis vectors modulated by matching coefficients:

$$\tilde{\mathbf{h}} = \sum_{k=1}^{K} \alpha_k \mathbf{e}_k \tag{5}$$

To enable dimension adaption between the latent composition space $\mathbb{R}^K$ and target feature space $\mathbb{R}^d$, we introduce a learnable projection operator:

$$\tilde{\mathbf{h}} = \mathbf{W}\left(\sum_{k=1}^{K} \alpha_k \mathbf{e}_k\right), \quad \mathbf{W} \in \mathbb{R}^{d \times K} \tag{6}$$

where $\mathbf{W}$ denotes a $d \times K$ trainable matrices. This parametric transformation enables: 1) dimension scaling between spaces, 2) learned feature recombination, and 3) differentiable composition through linear operator theory.

3.4. Category Prototype Matching

The final prediction is obtained through prototype-based similarity analysis. Let $\mathcal{C} = \{\mathbf{c}_y\}_{y=1}^{C}$ denote the learnable category prototypes in the latent space, where each $\mathbf{c}_y \in \mathbb{R}^n$ corresponds to a

 doi:10.20944/preprints202504.0003.v1

specific disease class, n is the oral disease classification number. The classification decision rule is formulated as:

$$\hat{y} = \underset{y \in \{1,\ldots,C\}}{\arg\max}\ \sigma\left(\frac{\langle \tilde{\mathbf{h}}, \mathbf{c}_y \rangle}{\tau}\right) \tag{7}$$

where σ denotes the softmax operator and τ represents the temperature hyperparameter.

Prototype Contrastive Loss: The model is optimized using a cross-entropy loss with prototype-based contrastive learning:

$$\mathcal{L}_{\text{con}} = -\mathbb{E}_{(\mathbf{x},y)}\left[\log \frac{\exp\left(\tau^{-1}\langle \tilde{\mathbf{h}}_x, \mathbf{c}_y \rangle\right)}{\sum_{j=1}^{C} \exp\left(\tau^{-1}\langle \tilde{\mathbf{h}}_x, \mathbf{c}_j \rangle\right)}\right] \tag{8}$$

where $\tilde{\mathbf{h}}_x$ represents the projected representation of input $\mathbf{x}$ using Eq. (6), and $\mathcal{L}_{\text{con}}$ is the contrastive loss. Note that the $\mathcal{L}_{\text{task}}$ defined in Eq. (1) is equal to $\mathcal{L}_{\text{con}}$.

4. Experiments

We conducted extensive experiments to validate the effectiveness of our proposed method. This section describes our experimental setup and results.

4.1. Experimental Setup

This part describes the experimental setups including dataset, implemental details, compared methods, and evaluation metrics.

4.1.1. Dataset

Our experiments utilize the **Dental Condition Dataset**, a clinically curated collection of dental images annotated for diagnostic and research applications. The dataset covers six dental diseases:

Caries: Images depicting tooth decay, cavities, or carious lesions.
Gingivitis: Cases of inflamed or infected gum tissue.
Hypodontia: Evidence of congenital or acquired absence of one or more teeth.
Mouth Ulcers: Visible canker sores or ulcerative lesions in oral mucosa.
Calculus: Examples of dental calculus or tartar buildup on tooth surfaces.
Tooth Discoloration: Manifestations of intrinsic or extrinsic tooth staining.

The entire dataset contains a total of 15,439 images, and all images are sourced from multisite hospital collaborations and public dental repositories. In our experiments, we selected 65% of the images for training, 20% for validation, and the remaining 15% for testing. During training, the dataset was augmented with rotation, horizontal flipping, scaling, and Gaussian noise injection. Some typical samples are shown in Figure 3.

 doi:10.20944/preprints202504.0003.v1

Figure 3. Typical image samples from the Dental Condition Dataset.

4.1.2. Implemental Details

In our implementation, we set the feature dimension d to 512, orthonormal bases count K to 512, temperature hyperparameter τ to 1, and the layer number in Equation (1) L to 12. All images were resized to 128×128 pixels for consistent network inputs. The model trained for 50 epochs with a batch size of 16, balancing computational demands and convergence quality. Experiments ran on PyTorch using an NVIDIA 3090 GPU with 24 GB memory. We utilized the AdamW optimizer with decoupled weight decay regularization and implemented OneCycleLR scheduling with a maximum learning rate of 0.01, which peaks after 30% of training before gradually decreasing according to per-epoch training steps.

4.1.3. Compared Methods

We compared our DyCoP model with several state-of-the-art computer vision architectures. The comparison models include: (1) classic CNN architectures (DenseNet121 [7], MobileNet-v2 [8], ResNet50 [10], VGG16 [11], VGG19 [11]); (2) EfficientNet-family models [41] (B0, B5); (3) EfficientViT variants [42] (B3, L3, M3); (4) Inception-family networks (InceptionResnet-v2 [43], Inception-v3 [44]); and (5) Transformer-based models (CrossViT [45], DEIT [46], TNTTransformer [47], Vision Transformer [48]). This diverse set of baseline models spans different architectural paradigms from traditional CNNs to modern transformer-based approaches.

4.1.4. Evaluation Metrics

To comprehensively evaluate the performance of our classification model, we employ multiple metrics including Accuracy, Macro-Precision, Macro-Recall, and Macro F_1-score.

 doi:10.20944/preprints202504.0003.v1

Accuracy measures the proportion of correctly classified instances among the total instances:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{9}$$

While accuracy provides an overall performance measure, it may be misleading for imbalanced datasets. Therefore, we also utilize class-specific metrics averaged across all classes:

Macro-Precision calculates the average precision across all classes, where precision for each class is defined as:

$$\text{Macro-Precision} = \frac{1}{C}\sum_{i=1}^{C}\text{Precision}_i, \quad \text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{10}$$

Macro-Recall similarly averages the recall across all classes:

$$\text{Macro-Recall} = \frac{1}{C}\sum_{i=1}^{C}\text{Recall}_i, \quad \text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{11}$$

Macro F1-score combines precision and recall into a single metric that balances both considerations:

$$\text{Macro F}_1\text{-score} = \frac{1}{C}\sum_{i=1}^{C}\text{F1-score}_i, \quad \text{F}_1\text{-score}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{12}$$

where C represents the number of classes, and TP_i, FP_i, and FN_i denote the true positives, false positives, and false negatives for class i, respectively.

The macro-averaging method treats all classes equally regardless of their size, making it particularly suitable for evaluating performance on imbalanced datasets where minority classes are as important as majority classes.

4.2. Experimental Results and Analysis

In this part, we list the quantitative experimental results and analyze the data to evaluate our proposed method DyCoP.

4.2.1. Performance Comparison with Other Models

As shown in Table 1, the proposed DyCoP model demonstrates superior performance compared to all other evaluated models. DyCoP achieves a classification accuracy of 93.3%, outperforming the second-best model, InceptionResnet-v2 (92.5%), by a margin of 0.8%. Furthermore, DyCoP attains the highest Macro-Precision of 92.5% and Macro-F1 Score of 91.9%, establishing its effectiveness in the classification of oral diseases.

Traditional CNN architectures, such as VGG16 and VGG19, exhibit substantially lower performance with accuracies of 64.5% and 50.3%, respectively. Notably, Transformer-based models, including Vision Transformer (75.6%) and CrossViT (79.5%), underperform compared to both our proposed model and other CNN-based architectures. This suggests that pure attention mechanisms may not sufficiently capture the intricate features associated with oral diseases without appropriate architectural modifications.

Among the evaluated models, EfficientNet-B0 (90.8%) and EfficientViT-B3 (91.0%) demonstrate competitive performance, indicating the potential of efficient architectures in this domain. However, the significant performance gap between these models and DyCoP highlights the effectiveness of our approach in addressing the challenges inherent in oral disease classification.

 doi:10.20944/preprints202504.0003.v1

Table 1. Performance comparison of our proposed DyCoP and other compared models. The best two methods for each metric are highlighted in **bold** and underlined, respectively.

Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F_1 Score
DenseNet121	89.6%	89.7%	89.6%	89.6%
MobileNet-v2	89.4%	89.8%	89.4%	89.6%
ResNet50	82.7%	82.8%	82.7%	82.2%
VGG16	64.5%	62.1%	64.5%	61.2%
VGG19	50.3%	48.9%	50.3%	44.2%
EfficientNet-B0	90.8%	90.9%	90.8%	90.4%
EfficientNet-B5	86.8%	87.5%	86.8%	86.9%
EfficientViT-B3	91.0%	89.5%	90.3%	89.8%
EfficientViT-L3	90.5%	89.5%	90.0%	89.7%
EfficientViT-M3	89.5%	89.1%	88.7%	88.8%
InceptionResnet-v2	92.5%	91.1%	**91.9%**	91.3%
Inception-v3	84.3%	84.7%	84.3%	84.5%
CrossViT	79.5%	78.8%	77.6%	78.1%
DEIT	87.3%	86.2%	85.2%	85.6%
TNTTransformer	76.4%	75.1%	73.3%	73.8%
Vision Transformer	75.6%	75.7%	74.6%	74.9%
DyCoP (Ours)	**93.3%**	**92.5%**	91.2%	**91.9%**

4.2.2. Confusion Matrix Analysis

The confusion matrix shown in Figure 4 provides detailed insights into the model's classification performance across six oral conditions: Calculus, Caries, Gingivitis, Ulcers, Tooth Discoloration, and Hypodontia.

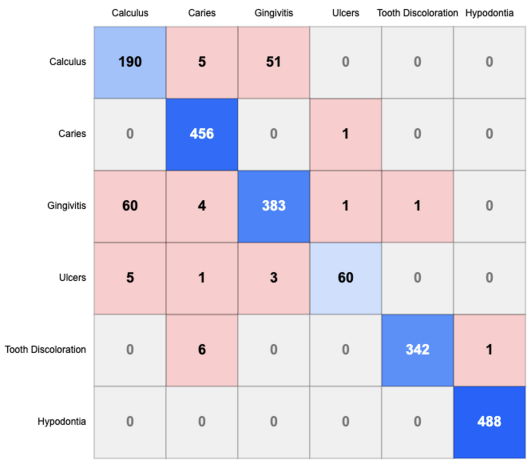

Figure 4. Visualization of the confusion matrix for the classification performance of the DyCoP model across six oral diseases (Calculus, Caries, Gingivitis, Ulcers, Tooth Discoloration, and Hypodontia).

Diagnostic Accuracy: The diagonal elements of the confusion matrix represent correctly classified instances, revealing strong performance across most categories. Hypodontia exhibits the highest number of correct classifications (488), followed by Caries (456) and Gingivitis (383). This demonstrates the model's robust ability to identify distinctive features associated with these conditions.

Cross-Misclassification Patterns: Several notable cross-misclassification patterns emerge from the confusion matrix:

- Calculus-Gingivitis Confusion: A substantial bidirectional misclassification exists between Calculus and Gingivitis, with 51 instances of Calculus misclassified as Gingivitis and 60 instances of Gingivitis misclassified as Calculus. This suggests morphological similarities between these conditions that challenge accurate differentiation.

 doi:10.20944/preprints202504.0003.v1

- Caries Classification: Caries demonstrates minimal confusion with other categories, indicating that its visual manifestations are sufficiently distinctive for accurate identification by the model.
- Tooth Discoloration: Six instances of Tooth Discoloration are misclassified as Caries, indicating potential visual similarities in certain presentations of these conditions.
- Hypodontia Recognition: Hypodontia shows negligible confusion with other conditions, achieving near-perfect classification with 488 correct identifications and no false negatives, suggesting highly distinctive visual characteristics.

Classification Robustness: The sparsity of certain regions in the confusion matrix indicates minimal confusion between specific condition pairs. For instance, Calculus and Tooth Discoloration show no mutual misclassification, as do Hypodontia and most other conditions. This demonstrates the model's capacity to distinguish between conditions with dissimilar visual presentations.

In summary, the experimental results validate the superior performance of our proposed DyCoP model across multiple evaluation metrics. The confusion matrix analysis reveals both the strengths of the model in differentiating most oral conditions and specific classification challenges, particularly between Calculus and Gingivitis. These insights provide valuable direction for future refinements, especially targeting improved discrimination between conditions that exhibit similar visual characteristics.

4.3. Convergence Analysis

Figure 5 presents the loss and accuracy curves for our model during training and validation. The loss curves demonstrate rapid convergence, with validation loss declining sharply from 1.63 to 0.5 within the first 4 epochs and ultimately stabilizing at 0.14. Similarly, the accuracy curves show swift improvement, with validation accuracy increasing from 32% to nearly 80% in early epochs before reaching a steady state of 92-93%. The consistent downward trend of the loss function and the corresponding increase in accuracy, both exhibiting stability in later epochs, provide compelling evidence that our proposed method converges efficiently and achieves robust performance on unseen data.

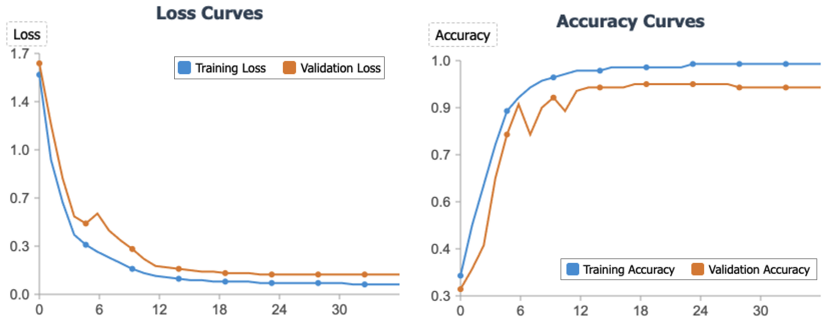

Figure 5. Training and Validation Curves.

4.4. Ablation Study

To validate the contribution of each component in our proposed DyCoP model, we conducted a comprehensive ablation study. Table 2 presents the quantitative results of this analysis, demonstrating the impact of various architectural choices on model performance.

 doi:10.20944/preprints202504.0003.v1

Table 2. Ablation study of DyCoP Components. We evaluate the impact of each component by removing it from the full model and the effect of varying the number of component prototypes. The best performance for each metric is highlighted in **bold**.

Variant	Accuracy	Macro-Precision	Macro-Recall	Macro-F_1 Score
DyCoP (Full)	**93.3%**	**92.5%**	**91.2%**	**91.9%**
DyCoP *w/o all designs* (ResNet-18)	83.3%	81.4%	85.0%	83.2%
DyCoP *w/o Componential Prototypes*	90.3%	89.7%	88.6%	89.1%
w/ 1024 (2×) Componential Prototypes	92.1%	91.5%	91.0%	91.3%
w/ 256 (half) Componential Prototypes	91.6%	90.9%	90.2%	90.5%
DyCoP *w/o Dynamic Comp. Rep.*	89.8%	88.9%	87.4%	88.1%
DyCoP *w/o Gradient Suppression*	91.1%	90.8%	89.5%	90.1%

4.4.1. Overall Performance Gain from Our Multiple Designs

To evaluate the cumulative effect of our proposed designs, we performed an ablation by removing all DyCoP-specific components, reverting to the baseline ResNet-18 architecture. As reported in Table 2, the ResNet-18 baseline achieves an accuracy of 83.3%, macro-precision of 81.4%, macro-recall of 85.0%, and macro-F_1 score of 83.2%. In contrast, the full DyCoP model attains a significantly higher accuracy of 93.3%, representing a performance gain of 10.0%. This substantial improvement highlights the critical role of our architectural enhancements in advancing the classification of oral diseases.

4.4.2. Effectiveness of Main Designs

We systematically ablated individual components from the DyCoP model to quantify their contributions to overall performance. The results are detailed below:

- **Gradient Suppression**: Eliminating the gradient suppression mechanism reduces accuracy by 2.2% (from 93.3% to 91.1%) and macro-F_1 score by 1.8% (from 91.9% to 90.1%). This decline underscores the importance of gradient suppression in mitigating overfitting and bolstering the model's generalization capacity.
- **Componential Prototypes**: Removing componential prototypes results in a notable decrease of 3.0% in accuracy (from 93.3% to 90.3%) and 2.8% in macro-F_1 score (from 91.9% to 89.1%). This indicates that componential prototypes are instrumental in capturing fine-grained features essential for distinguishing diverse oral disease manifestations. Furthermore, we investigated the impact of varying the number of componential prototypes on model performance:
  - **Increased Prototype Number**: Doubling the number of componential prototypes to 1024 yields a slight performance reduction compared to the default configuration, with accuracy decreasing by 1.2% (from 93.3% to 92.1%) and macro-F_1 score by 0.6% (from 91.9% to 91.3%). This suggests that an excess of prototypes may introduce redundancy, potentially capturing noise rather than meaningful patterns.
  - **Reduced Prototype Number**: Halving the number of componential prototypes to 256 results in an accuracy drop of 1.7% (from 93.3% to 91.6%) and a macro-F_1 score reduction of 1.4% (from 91.9% to 90.5%). This indicates that a minimum threshold of prototypes is required to adequately represent the diversity of features necessary for precise oral disease classification.
- **Dynamic Compositional Representation**: The absence of dynamic compositional representation leads to the most pronounced performance drop, with accuracy decreasing by 3.5% (from 93.3% to 89.8%) and macro-F_1 score by 3.8% (from 91.9% to 88.1%). This substantial degradation emphasizes the pivotal role of dynamic representations in generating discriminative features that effectively differentiate between oral disease categories.

The ablation study collectively substantiates the efficacy of our DyCoP architecture. Each component—gradient suppression, componential prototypes, and dynamic component representations—contributes significantly to the model's performance, with dynamic component representations

 doi:10.20944/preprints202504.0003.v1

exhibiting the greatest individual impact. Additionally, the analysis reveals that 512 componential prototypes strike an optimal balance between feature granularity and computational complexity, maximizing classification accuracy for oral disease diagnosis.

5. Conclusions

In this paper, we proposed a novel DyCoP framework for oral disease recognition, addressing critical challenges in automated dental diagnostics. Our method employs a novel architecture that disentangles disease representations into reusable componential prototypes and dynamically assembles them via sparse activation mechanisms. DyCoP includes several key technical designs: (1) a gradient suppression strategy that preserves hierarchical feature representations while adapting to domain-specific patterns; (2) componential prototype matching that identifies atomic visual concepts present in oral disease images; (3) dynamic compositional representation generation that adaptively combines these prototypes; and (4) category prototype matching that enables effective classification through prototype-based similarity analysis. Extensive experiments on the Dental Condition Dataset demonstrate the superior performance of our approach, and the ablation studies confirm the significant contribution of each component.

Author Contributions: Wenyuan Wang: Conceptualization, Methodology, Software, Investigation, Validation, Writing—original draft preparation; **Ruisi Yang**: Conceptualization, Investigation, Resources, Data curation, Writing—original draft preparation; **Jutao Xiao**: Formal analysis, Investigation, Writing—original draft preparation, Visualization; **Shuwei Huo**: Writing—original draft preparation, Writing—review and editing, Visualization, Supervision; **Zhengze Chen**: Writing—review and editing.

References

1. Zhao, Y.; Zhang, L.; Liu, Y.; Meng, D.; Cui, Z.; Gao, C.; Gao, X.; Lian, C.; Shen, D. Two-stream graph convolutional network for intra-oral scanner image segmentation. *IEEE Transactions on Medical Imaging* **2021**, *41*, 826–835.
2. Dixit, S.; Kumar, A.; Srinivasan, K. A current review of machine learning and deep learning models in oral cancer diagnosis: Recent technologies, open challenges, and future research directions. *Diagnostics* **2023**, *13*, 1353.
3. López-Cortés, X.A.; Matamala, F.; Venegas, B.; Rivera, C. Machine-learning applications in oral cancer: A systematic review. *Applied Sciences* **2022**, *12*, 5715.
4. Rashid, J.; Qaisar, B.; Faheem, M.; Akram, A.; Amin, R.; Hamid, M. Mouth and oral disease classification using InceptionResNetV2 method. *Multimedia Tools and Applications* **2024**, *83*, 33903–33921.
5. Can, Z.; Isik, S.; Anagun, Y. CVApool: using null-space of CNN weights for the tooth disease classification. *Neural Computing and Applications* **2024**, *36*, 16567–16579.
6. Kang, J.; Le, V.; Lee, D.; Kim, S. Diagnosing oral and maxillofacial diseases using deep learning. *Scientific Reports* **2024**, *14*, 2497.
7. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269.
8. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
9. Huo, S.; Zhou, Y.; Chen, K.; Xiang, W. Skim-and-scan transformer: A new transformer-inspired architecture for video-query based video moment retrieval. *Expert Systems with Applications* **2025**, *270*, 126525.
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
11. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the Proceedings of the International Conference on Learning Representations (ICLR), 2015.
12. Zhou, Y.; Wang, H.; Huo, S.; Wang, B. Hierarchical full-attention neural architecture search based on search space compression. *Knowledge-Based Systems* **2023**, *269*, 110507.

 doi:10.20944/preprints202504.0003.v1

13. Zhou, Y.; Hao, J.; Huo, S.; Wang, B.; Ge, L.; Kung, S.Y. Automatic metric search for few-shot learning. *IEEE Transactions on Neural Networks and Learning Systems* **2023**.
14. Adeoye, J.; Su, Y. Artificial intelligence in salivary biomarker discovery and validation for oral diseases. *Oral Diseases* **2024**, *30*, 23–37.
15. Zhou, M.; Jie, W.; Tang, F.; Zhang, S.; Mao, Q.; et al., C.L. Deep learning algorithms for classification and detection of recurrent aphthous ulcerations using oral clinical photographic images. *Journal of Dental Sciences* **2024**, *19*, 254–260.
16. Mira, E.; Sapri, A.; Aljehanı, R.; Jambı, B.; Bashir, T.; et al., E.E.K. Early diagnosis of oral cancer using image processing and artificial intelligence. *Fusion: Practice and Applications* **2024**, *14*, 293–308.
17. Babu, P.; Rai, A.; Ramesh, J.; Nithyasri, A.; Sangeetha, S.; et al., P.K. An explainable deep learning approach for oral cancer detection. *Journal of Electrical Engineering and Technology* **2024**, *19*, 1837–1848.
18. Rokhshad, R.; Mohammad-Rahimi, H.; Price, J.; Shoorgashti, R.; Abbasiparashkouh, Z.; et al., M.E. Artificial intelligence for classification and detection of oral mucosa lesions on photographs: a systematic review and meta-analysis. *Clinical Oral Investigations* **2024**, *28*, 88.
19. Zanini, L.; Rubira-Bullen, I.; Nunes, F. A systematic review on caries detection classification and segmentation from x-ray images: methods datasets evaluation and open opportunities. *Journal of Imaging Informatics in Medicine* **2024**, pp. 1–22.
20. Montagnoli, D.; Leite, V.; Godoy, Y.; Lafetá, V.; Junior, E.; et al., A.C. Can predictive factors determine the time to treatment initiation for oral and oropharyngeal cancer? A classification and regression tree analysis. *Plos One* **2024**, *19*, e0302370.
21. Deo, B.; Pal, M.; Panigrahi, P.; Pradhan, A. An ensemble deep learning model with empirical wavelet transform feature for oral cancer histopathological image classification. *International Journal of Data Science and Analytics* **2024**, pp. 1–18.
22. Ansari, A.; Singh, A.; Singh, M.; Kukreja, V. Enhancing Skin Disease Classification: A Hybrid CNN-SVM Model Approach. In Proceedings of the Proceedings of the 2024 International Conference on Automation and Computation (AUTOCOM), March 2024, pp. 29–32.
23. Holtfreter, B.; Kuhr, K.; Tonetti, M.; Sanz, M.; et al., K.K. ACES: A new framework for the application of the 2018 periodontal status classification scheme to epidemiological survey data. *Journal of Clinical Periodontology* **2024**, *51*, 512–521.
24. Liu, C.; Eim, I.; Kim, J. High accuracy handwritten Chinese character recognition by improved feature matching method. In Proceedings of the Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR). IEEE, 1997, Vol. 2, pp. 1033–1037.
25. Yang, H.; Zhang, X.; Yin, F.; Liu, C. Robust classification with convolutional prototype learning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3474–3482.
26. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4077–4087.
27. Chen, Y.; Zhu, X.; Gong, S. Semi-supervised deep learning with memory. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 268–283.
28. Han, J.; Luo, P.; Wang, X. Deep self-learning from noisy labels. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 5138–5147.
29. Kuo, C.; Ma, C.; Huang, J.; Kira, Z. Manifold graph with learned prototypes for semi-supervised image classification. *arXiv preprint* **2019**, *arXiv:1906.05202*.
30. Zhou, L.; Chen, H.; Wei, Y.; Li, X. Mining confident supervision by prototypes discovering and annotation selection for weakly supervised semantic segmentation. *Neurocomputing* **2022**, *501*, 420–435.
31. Zhou, T.; Zhang, M.; Zhao, F.; Li, J. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4299–4309.
32. Yang, B.; Liu, C.; Li, B.; Jiao, J.; Ye, Q. Prototype mixture models for few-shot semantic segmentation. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2020, pp. 763–778.
33. Du, Y.; Fu, Z.; Liu, Q.; Wang, Y. Weakly Supervised Semantic Segmentation by Pixel-to-Prototype Contrast. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4320–4329.

 doi:10.20944/preprints202504.0003.v1

34. Chen, Q.; Yang, L.; Lai, J.; Xie, X. Self-supervised Image-specific Prototype Exploration for Weakly Supervised Semantic Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4288–4298.
35. Cheng, P.; Lin, L.; Lyu, J.; Huang, Y.; Luo, W.; Tang, X. Prior: Prototype representation joint learning from medical images and reports. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2023, pp. 21361–21371.
36. Song, G.; Liu, Y.; Wang, X. Revisiting the Sibling Head in Object Detector. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11563–11572.
37. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 658–666.
38. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2020, pp. 213–229.
39. Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; Ling, H. Lasot: A high-quality benchmark for large-scale single object tracking. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5374–5383.
40. Huang, L.; Zhao, X.; Huang, K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2019**, *43*, 1562–1577.
41. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the Proceedings of the International Conference on Machine Learning (ICML), 2019.
42. Cai, H.; Li, J.; Hu, M.; Gan, C.; Han, S. EfficientViT: Lightweight Multi-Scale Attention for High-Resolution Dense Prediction. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2023.
43. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2017.
44. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception architecture for computer vision. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.
45. Chen, C.F.R.; Zhu, Y.; Sukthankar, R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.
46. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers and distillation through attention. In Proceedings of the Proceedings of the International Conference on Machine Learning (ICML), 2021.
47. Han, K.; Xiao, A.; Wu, E.; et al.. Transformer in Transformer. *arXiv preprint arXiv:2103.00112* **2021**.
48. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al.. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the Proceedings of the International Conference on Learning Representations (ICLR), 2021.