Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 21(2): 2366–2384.
DOI: 10.3934/mbe.2024104
Received: 16 November 2023
Revised: 22 December 2023
Accepted: 23 December 2023
Published: 15 January 2024

Research article

STS-TransUNet: Semi-supervised Tooth Segmentation Transformer U-Net for dental panoramic image

Duolin Sun[1,2,†], Jianqing Wang[3,†], Zhaoyu Zuo[1], Yixiong Jia[4] and Yimou Wang[1,*]

[1] University of Science and Technology of China, Hefei, China
[2] Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China
[3] Hangzhou Sai Future Technology Co., Ltd, Hangzhou, China
[4] Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China

† **Co-First Authors:** These two authors contributed equally.

* **Correspondence:** Email: wangyimou@mail.ustc.edu.cn.

Abstract: In this paper, we introduce a novel deep learning method for dental panoramic image segmentation, which is crucial in oral medicine and orthodontics for accurate diagnosis and treatment planning. Traditional methods often fail to effectively combine global and local context, and struggle with unlabeled data, limiting performance in varied clinical settings. We address these issues with an advanced TransUNet architecture, enhancing feature retention and utilization by connecting the input and output layers directly. Our architecture further employs spatial and channel attention mechanisms in the decoder segments for targeted region focus, and deep supervision techniques to overcome the vanishing gradient problem for more efficient training. Additionally, our network includes a self-learning algorithm using unlabeled data, boosting generalization capabilities. Named the Semi-supervised Tooth Segmentation Transformer U-Net (STS-TransUNet), our method demonstrated superior performance on the MICCAI STS-2D dataset, proving its effectiveness and robustness in tooth segmentation tasks.

Keywords: dental panoramic image; attention mechanisms; STS-TransUNet; self-training; semi-supervised

1. Introduction

Oral health is a pivotal aspect of overall well-being, with dental ailments such as periodontal disease, cavities, and misalignments not only affecting masticatory function and aesthetics but also

potentially correlating with systemic maladies like cardiovascular diseases and diabetes [1]. In the field of dental diagnostics, panoramic imaging, also known as orthopantomography, has become increasingly significant [2]. This technology provides a comprehensive view of the mouth, capturing images of all the teeth and the surrounding bone structure in a single shot. Unlike the traditional intraoral radiography, panoramic imaging offers a broad perspective, essential for a holistic assessment of dental health. It is particularly invaluable in identifying problems in areas such as tooth positioning, impacted teeth, and the development of tumors [3–6]. Moreover, in orthodontic treatments, tooth extractions, and pre-surgical planning, panoramic images offer clinicians a clear and detailed view, crucial for designing precise orthodontic appliances, assessing surgical risks, and formulating effective treatment plans, thereby significantly enhancing patient care [7, 8]. Tooth segmentation not only significantly reduces diagnostic time and enhances diagnostic accuracy but also furnishes vital information for pathological analysis and personalized treatment planning [6]. For instance, accurate tooth segmentation can aid in evaluating the relationship between teeth and alveolar bone, determining the optimal position for dental implants, or assessing the outcomes of orthognathic surgery [9]. However, manual tooth segmentation in panoramic imaging interpretation, a task for radiologists and dental specialists, is time-consuming and costly, underscoring the urgent clinical need for automated segmentation technology to assist medical professionals in efficient and accurate diagnostics.

In recent years, the medical imaging field has witnessed a significant transformation with the rapid development of deep learning [10–12]. Unlike traditional methods that rely on manual feature extraction [13], deep learning can identify and categorize the complex and diverse features of both the 1D physilogical parameters and 2D medical images [14–16]. The capability of deep learning for automatic feature extraction in medical imaging leads to the creation of robust, quantifiable models with strong adaptability and generalizability, significantly aiding doctors in formulating precise and effective medical plans [17–19]. The advent of automatic tooth segmentation technologies [20], leveraging and computer vision techniques, has the potential to autonomously identify and segment dental structures [21, 22].

Current approaches predominantly utilize U-shaped convolutional neural network architectures, with methods like Faster R-CNN [23] and Mask R-CNN [24] being widely applied in tooth segmentation and caries detection [25, 26]. However, these are typically only suitable for downsampled Cone Beam Computed Tomography (CBCT) images. MSLPNet [27] employs a multi-scale structure to mitigate boundary prediction issues, subsequently utilizing a location-aware approach to pinpoint each dental pixel in panoramic images. Finally, an aggregation module is incorporated to diminish the semantic discrepancies across multiple branches. Two-stage segmentation methods [28, 29] generally locate the approximate position of the teeth in the first phase, followed by precise segmentation in the second. In a similar vein, the model in [30] introduces a coarse-to-fine tooth segmentation strategy, pre-trained on large-scale, weakly supervised datasets to initially locate teeth, and then fine-tuned on smaller, meticulously annotated datasets. Beyond weak supervision, researchers often resort to semi-supervised learning strategies with limited annotated data, such as self-training and pseudo-label generation. A novel semi-supervised 3D dental arch segmentation pipeline is proposed by [31], utilizing k-means for self-supervised learning [32, 33] and supervised learning on annotated data. The pipeline in [34] refines nnU-Net [35] architecture, training a preliminary nnU-Net model and then allowing medical professionals to supervise its performance

on unannotated datasets, selectively updating the model. Undoubtedly, this semi-supervised approach is cost-intensive. Overall, while these methods have achieved commendable performance, convolution-based approaches are limited by their receptive field for relatively larger input images and rely on prior localization of teeth.

The long-range dependency capabilities of Transformer architectures [36–39] have inspired new paradigms in image processing. The sequence attention mechanism of vision Transformers aggregates different patches of the same image, allowing each patch to interact with others, a significant advantage over CNNs with their inductive bias priors. Transformers have similarly revolutionized medical imaging [40]. TransUNet [41] introduces a U-Net combined with Transformer architecture for medical segmentation, merging CNN's local focus with the Transformer's global feature extraction capabilities, significantly inspiring the medical segmentation field. BoTNet [42], blending Transformers with convolutions, proposes a lightweight instance segmentation backbone, replacing some of the final convolutional layers of ResNet with Transformers. Building on this, GT U-Net [43] introduces a Fourier loss leveraging dental prior knowledge, effectively segmenting dental roots. However, while these Transformer-based methods excel in capturing global interactions in the encoder, they often do not optimally leverage these encoded features due to limitations in their decoding mechanisms. This leads to certain deficiencies in current deep learning approaches to tooth segmentation, resulting in suboptimal performance.

To this end, we introduce STS-TransUNet, a model that merges a CNN-Transformer encoder—blending CNN's shallow local feature extraction with Transformer's deep global encoding [44] and a customized upsampling module as the decoder—aiming at prioritizing key information and filtering out the redundant. Specifically, we have innovated the decoder part of the architecture by incorporating channel and spatial attention mechanisms [45]. This enables the decoder to focus exclusively on pertinent information while disregarding redundant data. The use of deep supervision techniques allows for immediate feedback on each layer of the decoder, thereby accelerating the convergence rate. By integrating the input and output images, our method enhances the model's ability to directly associate and learn from the initial and desired final states of the images. This novel strategy overcomes some of the limitations observed in traditional segmentation methods, where a disconnect between input and processed images can lead to inefficiencies and inaccuracies. Furthermore, we employ a straightforward self-training semi-supervised strategy, effectively segmenting the MICCAI 2023 public challenge dataset (STS-2D) [46] and achieving a distinguished position in the competition. The primary contributions of this paper are threefold:

1). We propose the STS-TransUNet, a novel single-stage model tailored for precise and automated segmentation in clinical dentistry. This model is specifically devised for panoramic dental imaging and leverages advanced deep learning techniques to accurately identify and outline dental structures.

2). A decoder with spatial and channel attention mechanisms, combined with deep supervision techniques, effectively captures the irregularities in dental information, mitigates gradient vanishing, and accelerates convergence.

3). Extensive experiments conducted on the MICCAI STS-2D dataset demonstrate the exemplary performance of our approach.

2. Related work

2.1. CNN-based methods in medical segmentation

Deep learning, particularly CNN-based approaches, has demonstrated exceptional performance across a broad spectrum of practical applications [47–53], including the domain of medical image segmentation. Diverging from traditional approaches in medical image segmentation [21, 54–56], the advent of the U-Net [10] architecture has heralded a new era in this field, significantly enhancing the precision and efficiency of segmentation tasks. Its encoder-decoder structure was capable of extracting high-level features from input images and using them to generate fine segmentation results [35]. In [57], deep learning methods were first introduced into panoramic X-ray tooth segmentation. Specifically, they performed pre-training on the backbone using the Mask R-CNN on the MSCOCO dataset and fine-tune it on their own dataset. In [58], the influence of factors such as data augmentation, loss functions, and network ensembles on tooth segmentation based on U-Net was investigated, fully exploiting the performance of the U-Net. TSegNet [59] formulated the 3D tooth point cloud data segmentation task as the precise localization of each tooth's center based on distance perception and the segmentation task based on confidence perception. This task was accomplished through accurate positioning in the first stage and precise segmentation in the second stage. All the aforementioned methods employed supervised deep learning techniques. In the realm of semi-supervised learning, MLUA [60] adopted a teacher-student strategy, utilizing a single U-shaped network for both annotated and unannotated data. Considering the irregular shape and significant variability of teeth, this model introduced multi-level perturbations to train more robust systems. Similarly, the model proposed in [61] employed a comparable strategy, focusing on data augmentation in areas of carious lesions, resulting in a high-performing caries segmentation model. The proposal in [34] relied on the expertise of medical professionals to select data for semi-supervised segmentation. The success of these methods largely hinged on the profound impact of CNNs in image processing. However, CNNs inherently possess inductive bias limitations, particularly in their local feature extraction. In contrast to the aforementioned methods, our approach integrates CNN's capability for shallow local feature extraction with global Transformer encoding, thereby achieving comprehensive global capture of dependencies.

2.2. Transformer-based methods in medical segmentation

CNN-based methods inherently possess inductive biases and struggle to effectively learn global semantic interactions due to the locality of the convolution operation [62]. TransUNet [41] pioneered a new paradigm in medical segmentation by integrating the global encoding capabilities of Transformers with the upsampling features of U-Net. Following this, a multitude of methods based on the TransUNet framework have been custom-tailored and applied to various other domains of medical image segmentation [63, 64], demonstrating its versatility and effectiveness. UNETR [65] took this further by transforming volumetric medical images into a sequence prediction problem, marking a significant application of Transformers in 3D medical imaging. Swin-Unet [66] merged the entire topological structure of Unet with the attention mechanisms of Swin Transformer. Its decoder used patch expanding for upsampling and showed remarkable performance on multi-organ CT and ACDC datasets. Similarly, the model in [67] developed a multi-task architecture based on Swin Transformer

for segmenting and identifying teeth and dental pulp calcification. The Mask-Transformer-based architecture [68] has demonstrated impressive capabilities in tooth segmentation. It employed a dual-path design combined with a panoramic quality loss function to simplify the training process. While these methods leveraged the global dependency capabilities of Transformer encoders, they often overly focused on global feature extraction by the encoder. Moreover, few studies have explored combining Transformer methods with actual unannotated dental panoramic image data segmentation. Unlike methods based on pure Transformer encoder-decoder architectures, our encoder employs a CNN-Transformer architecture, maximizing the use of U-Net's skip connections. This design choice is informed by the inherent limitation of Transformer architectures in not effectively capturing global dependencies at shallower layers [44, 69]. Our decoder focuses on relevant information without the need for prior tooth localization, employing a straightforward self-training method to generate pseudo-labels and iteratively update the model. This approach has demonstrated excellent performance on the MICCAI STS-2D dataset [46].

3. Materials and methods

3.1. Materials

We utilize a high-quality MICCAI STS-2D dataset [46], including panoramic dental CT images of children aged 3–12 years, obtained from Hangzhou Dental Group, Hangzhou Qiantang Dental Hospital, Electronic Science and Technology University, and Queen Mary University of London. The dataset, serving as the official training set, comprises a total of 5000 images, including 2900 labeled and 2100 unlabeled images. All our experiments utilize this training set as the primary dataset. Our model's results on the official test set are detailed in Section 4.3.

3.2. Data partition and preprocessing

Data split: Fully supervised training data (random 2500 labeled images) are employed for fully supervised training. Semi-supervised training data (random 2000 labeled images and 2100 unlabeled images) are used for semi-supervised training. Test data (400 and 900 labeled images) are reserved for testing fully supervised and semi-supervised method, respectively.

Data preprocessing: The original images have a resolution of 640×320. To facilitate training, we resize them to 640×640, then further downsample them to 320×320 and 160×160. These smaller sizes are used for deep supervised training. During training, we apply data augmentation strategies, including random flips, rotations, and cropping, to enhance model robustness and performance.

3.3. Network architecture

In our approach, we adopt the well-established Unet architecture, which comprises two fundamental components: An encoder and a decoder, as shown in Figure 1. The encoder plays a crucial role in extracting high-level features from the input images, while the decoder is responsible for generating the final segmented results. We represent our model with the following formulas:

$$H = \text{ViT}(\text{LinearProjection}(\text{ResNet50}(X))), \tag{3.1}$$

$$O_1, O_2, O_3 = \text{CNNDecoder}(H), \tag{3.2}$$

where H denotes the hidden feature obtained from the CNN-Transformer hybrid encoder, and O_1, O_2, O_3 represent the outputs from the last three layers of the CNN decoder, which are used for deep supervised training.

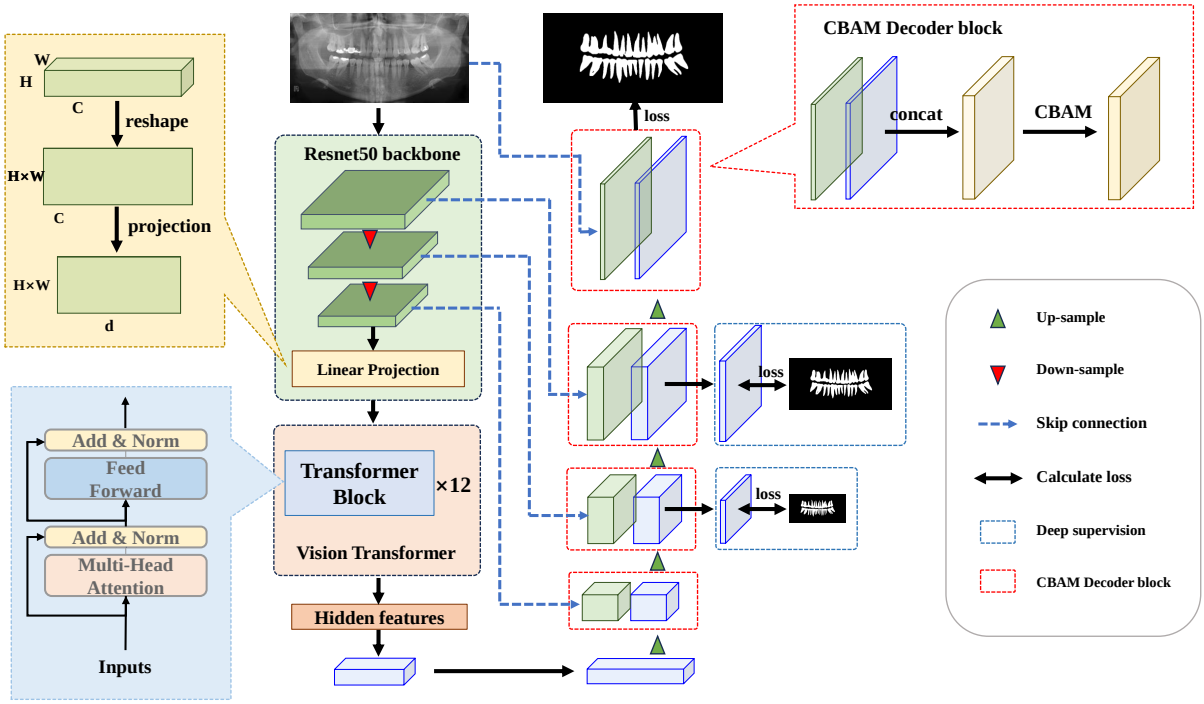

Figure 1. The architecture of our proposed STS-TransUNet. ResNet50 backbone is the standard and classical CNN backbone. The module following ResNet50 is ViT with 12 Transformer blocks. ResNet50 is employed for local short-range contextual modeling and ViT is for global long-range contextual modeling. CBAM is used for spatial and channel features aggregation. Outputs of the last three layers are used for calculating loss, that is known as deep supervision.

Recognizing the unique strengths of both convolutional neural networks (CNNs) and Transformers, we design a hybrid encoder structure. CNNs excel at capturing position-aware features, while Transformers are proficient at integrating long-range contextual information. By combining these two architectural elements, we harness their complementary advantages. This hybrid encoder structure enhances the model's ability to comprehend the underlying content within the images.

For the decoder, we employ a standalone CNN architecture. This choice aims to facilitate the model's effective learning of spatial and channel-related information. To further enhance performance, we introduce the Convolutional Block Attention Module (CBAM) [45].

CBAM is an attention mechanism employed in computer vision tasks with the primary objective of enhancing the performance of convolutional neural networks (CNNs). It enables better focus on important information in different channels and spatial locations when processing images. CBAM consists of two key components: Channel attention and spatial attention. Channel attention helps the model learn which channels are most crucial for tasks such as image classification, while spatial

attention helps the model identify essential regions or positions in an image for the task. This adaptive weighting mechanism allows the model to adapt to various images and tasks. Moreover, CBAM has demonstrated significant performance improvements in computer vision tasks, including image classification, object detection, and semantic segmentation. Its main advantage lies in its ability to automatically learn which features are more important for a given task, thus enhancing the model's performance and robustness. CBAM has found wide application in deep learning, providing a potent tool for the field of computer vision.

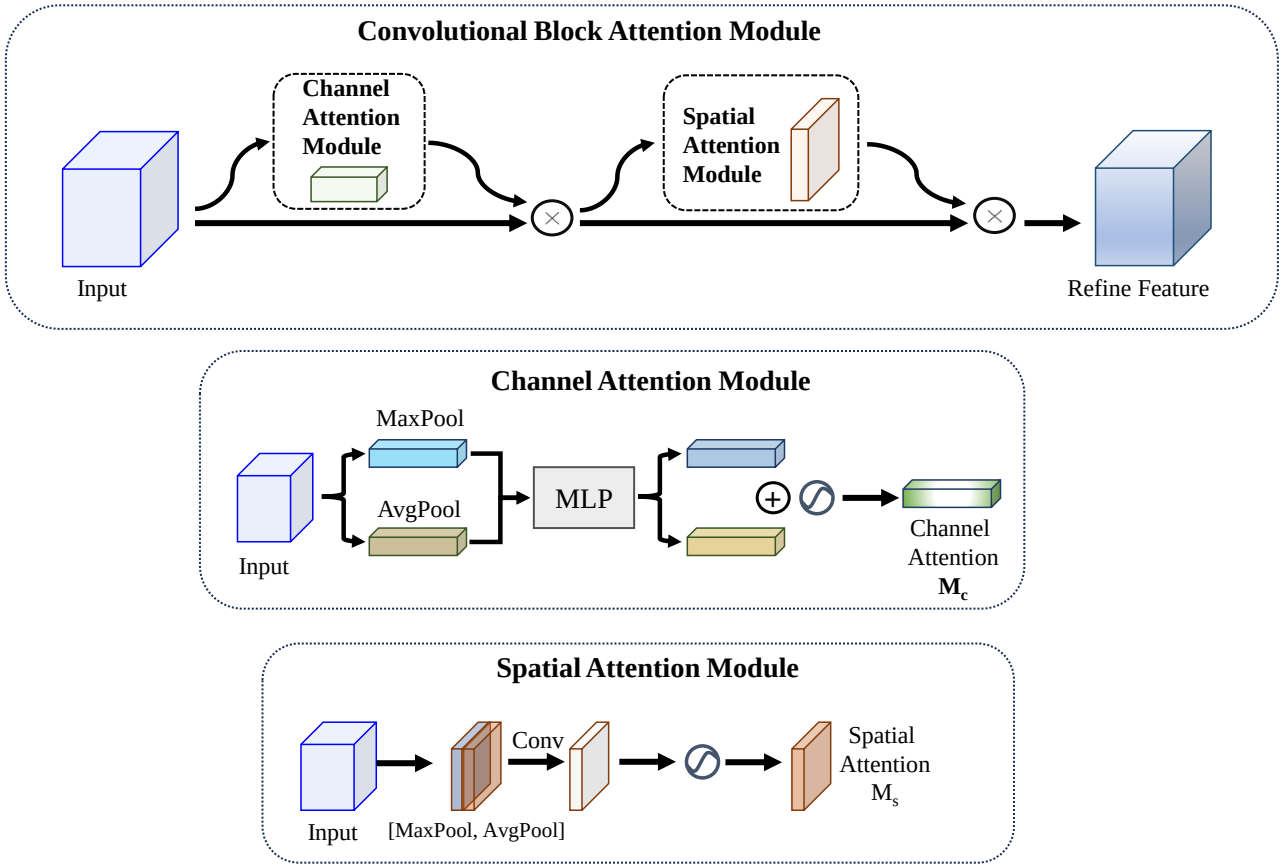

Figure 2. The architecture of CBAM. CBAM is composed of CAM and SAM. That means CBAM provides a comprehensive attention mechanism, improving a model's ability to capture meaningful patterns.

$$F' = M_c(F) \otimes F, \tag{3.3}$$

$$F'' = M_s(F') \otimes F', \tag{3.4}$$

where $\otimes$ denotes element-wise multiplication. During multiplication, the attention values are broadcasted (copied) accordingly: channel attention values are broadcasted along the spatial dimension, and vice versa. F'' is the final refined output. Figure 2 depicts the computation process of each attention map. Following formulas describe the details of each attention module:

$$
\begin{aligned}
M_c &= \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \\
&= \sigma(W_1(W_0(F_{c_{\text{avg}}})) + W_1(W_0(F_{c_{\text{max}}}))), \qquad (3.5) \\
M_s &= \sigma(f_{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])) \\
&= \sigma(f_{7\times7}([F_{s_{\text{avg}}}; F_{s_{\text{max}}}])), \qquad (3.6)
\end{aligned}
$$

The CBAM module enhances proposed model's understanding of image content and assists it in prioritizing specific channels. To expedite the model's convergence during training and enhance its ability to generalize to different image scales, we implement a deep supervised training strategy. This strategy involves introducing supervised signals into the last three decoder layers, each corresponding to different image scales. It enables the model to better understand and adapt to various image scales effectively.

3.4. Training strategy

We train both fully supervised and semi-supervised models using a loss function that combines dice loss and IoU loss weighting. The formula of used total loss is as following:

$$loss = DeepDiceLoss(\hat{Y}_{deep}, Y_{deep}) \times 0.6 + DeepIoULoss(\hat{Y}_{deep}, Y_{deep}) \times 0.4, \tag{3.7}$$

$$DiceLoss = 1 - \frac{2 \cdot |\hat{Y} \cap Y|}{|\hat{Y}| + |Y|}, \tag{3.8}$$

$$IoULoss = 1 - \frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|}, \tag{3.9}$$

where $\hat{Y}$ and Y respectively represent the prediction and ground truth. We employ deep supervision by computing the loss at three different scales, the deep supervision loss are shown as following:

$$DeepLoss = loss(\hat{Y}_{640}, Y_{640}) + loss(\hat{Y}_{320}, Y_{320}) + loss(\hat{Y}_{160}, Y_{160}). \tag{3.10}$$

where $\hat{Y}_{640}, \hat{Y}_{320}, \hat{Y}_{160}$ denote the outputs of the last three decoder layers, each with resolutions of 640, 320, and 160, respectively. Similarly, $Y_{640}, Y_{320}, Y_{160}$ represent the ground truth, resized to correspond to these resolutions. The following section details the specific two-stage training process.

3.4.1. First stage: Fully supervised training

In this stage, we implement a fully supervised training approach using samples with real labels, adopting a 5-fold cross-validation strategy to enhance model robustness. Instead of treating each fold's output as a separate model, we integrated these models from all five folds into a single ensemble model. This ensemble approach capitalizes on the strengths of each fold's training, resulting in a more robust and generalized model that effectively captures the diversity of the training data. The details of the first-stage training are as follows:

Learning rate initialization: We set the initial learning rate to 3e-4 to ensure a stable start to the training process.

Total training epochs: The training process encompasses a total of 200 epochs, providing the model with sufficient time to progressively enhance its performance. However, it is common for the initial few epochs to exhibit some instability.

Warm-up strategy: To mitigate the model's instability at the beginning of training, we implement a warm-up strategy. This involves gradually increasing the learning rate within the first 3 epochs, guiding the model towards a more stable training state.

Cosine curve strategy: Subsequent adjustments to the learning rate follow a cosine curve strategy. This strategy gradually reduces the learning rate, allowing for a more refined adjustment of model parameters until the learning rate decays to 0. This aids the model in better convergence during the later stages of training.

Fully supervised training is conducted to establish the foundational performance of the model, enabling it to learn feature extraction from labeled data and perform tasks. This training phase equips the model with a certain degree of predictive capability, laying the groundwork for subsequent semi-supervised learning.

3.4.2. Second stage: Semi-supervised training

In the second stage, we employ a semi-supervised training approach, capitalizing on the benefits it offers. Specifically, we adopt the self-training strategy [70] to generate pseudo-labels and facilitate model training. This phase harnesses unlabeled data effectively, maximizing the utility of available resources. The workflow of used self-training is shown as below:

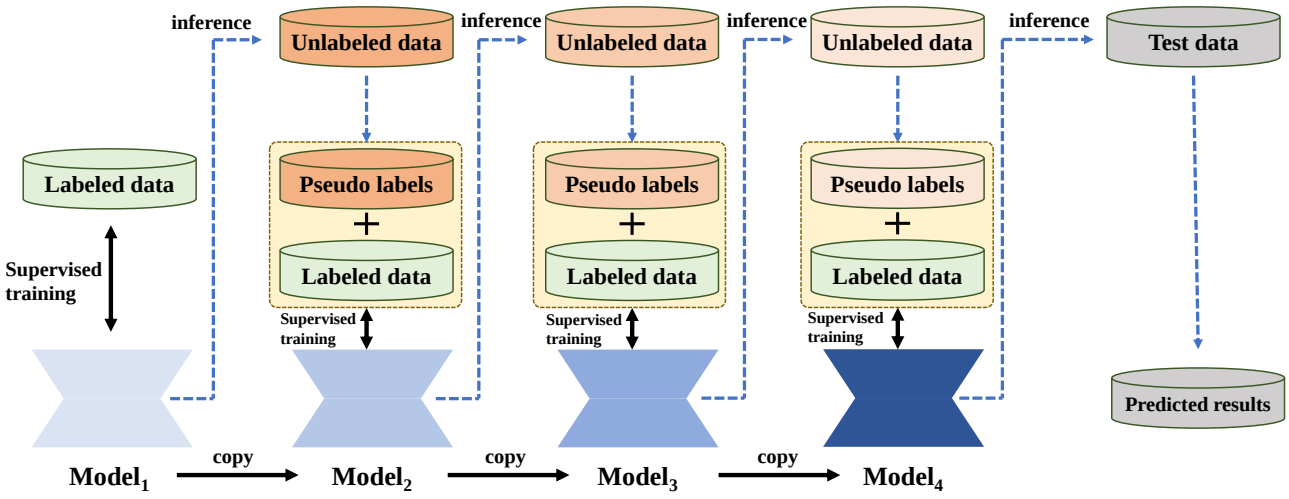

Figure 3. The workflow of self-training. Labeled data are used for supervised training, unlabeled data will be used to generate pseudo labels and add into labeled data for supervised training.

Generating Pseudo-Labels: We initiate this phase by feeding 2100 unlabeled images into the model obtained from the first supervised training stage. The model, in response, produces outputs containing predicted class (foreground and background) probabilities for these images. These

probabilities are then averaged across the entire set.

Pseudo-Label selection: To identify high-quality data points for training, we select the top 300 images based on the predicted probabilities. These images are paired with the corresponding high-quality pseudo-labels generated by the model.

Training with augmented data: The chosen images, along with their newly created pseudo-labels, are used to augment the training dataset. The training process initializes with an initial learning rate of 1e-4 and spanned three epochs. This helps the model adapt to the augmented dataset.

Iterative refinement: In pursuit of further model improvement, this process is repeated five times. In each iteration, a new model is employed, and the same steps are repeated. This iterative refinement strategy allows the model to learn progressively from the unlabeled data. This semi-supervised training strategy, specifically the self-training method, is valuable for harnessing the potential of unlabeled data, effectively expanding the training dataset, and improving the model's performance. It is a powerful tool for leveraging available resources and enhancing the robustness of the final model.

4. Experiments and analysis

All our experiments are conducted on two 32 GB V100 GPUs. Our STS-TransUNet has a training duration of 12 hours, and we use Pytorch 1.12 as the experimental framework. Additionally, to ensure the reproducibility of our results, we have fixed the seed in all our experiments.

4.1. Comparative results

Quantitative analysis: After fully supervised and semi-supervised training, the results are presented in Table 1. We use Dice, IoU (Intersection over Union) and Hausdorff distance as our evaluation metrics. The formulas of them are shown as below,

$$Dice = \frac{2 \cdot |\hat{Y} \cap Y|}{|\hat{Y}| + |Y|}, \tag{4.1}$$

$$IoU = \frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|}, \tag{4.2}$$

$$H(\hat{Y}, Y) = \max\left(\sup_{\hat{y}\in\hat{Y}} \inf_{y\in Y} d(\hat{y}, y), \sup_{y\in Y} \inf_{\hat{y}\in\hat{Y}} d(y, \hat{y})\right). \tag{4.3}$$

where $\hat{Y}$ and Y represent the prediction and the ground truth, respectively.

For the fully supervised training, models like UNet++ [71], UNet 3+ [72], and R2AU-Net [73], which rely solely on CNN, exhibits relatively weak perception of global information, resulting in less-than-ideal performance. Among the CNN-based models, R2AU-Net, which incorporates attention mechanisms, performed the best. While Transformer blocks are capable of capturing long-range information, they exhibit poorer position awareness inherently and require substantial training data to excel. As a result, the performance of Swin-Unet [66, 74], DAE-Former [75] and SegFormer [76] are not on par with our model. In summary, the hybrid combination of CNN and Transformer in our model harnesses the strengths of both and delivered satisfying results. Even in the semi-supervised training phase, our model outperforms the other models. While all models

experience a decrease in dice scores on the semi-supervised test set due to its larger size, our model retains its superior performance.

Table 1. Comparison of quantitative results on test data. Bold indicate the best results.

	Fully supervised			Semi-supervised		
	Dice	IoU	Hausdorff distance	Dice	IoU	Hausdorff distance
UNet++ [71]	0.8978	0.9560	0.0368	0.8689	0.9427	0.0403
UNet 3+ [72]	0.9070	0.9589	0.0326	0.8739	0.9531	0.0365
R2AU-Net [73]	0.9081	0.9598	0.0309	0.8826	0.9556	0.0321
SegFormer [76]	0.9182	0.9626	0.0304	0.9087	0.9589	0.0303
Swin-Unet [66]	0.9171	0.9631	0.0303	0.9102	0.9588	0.0301
DAE-Former [75]	0.9251	0.9685	0.0306	0.9153	0.9601	0.0286
STS-TransUNet (Ours)	**0.9318**	**0.9691**	**0.0298**	**0.9206**	**0.9723**	**0.0269**

Qualitative analysis: Results from different models on randomly selected 4 samples are presented in Figure 4. Comparing models solely based on CNN with those incorporating attention mechanisms, the latter achieves clearer results. However, in comparison to Transformer-based models, the ability to segment the completeness of teeth remains a challenge, affirming the notion that Transformers possess stronger global modeling capabilities relative to CNN. Nevertheless, models based exclusively on Transformers often struggle with local information awareness compared to CNN. This is evident in Figure 4, where DAE-Former, while superior in overall results to CNN models, falls slightly short in fine details. Our model outperforms others in terms of texture and completeness.

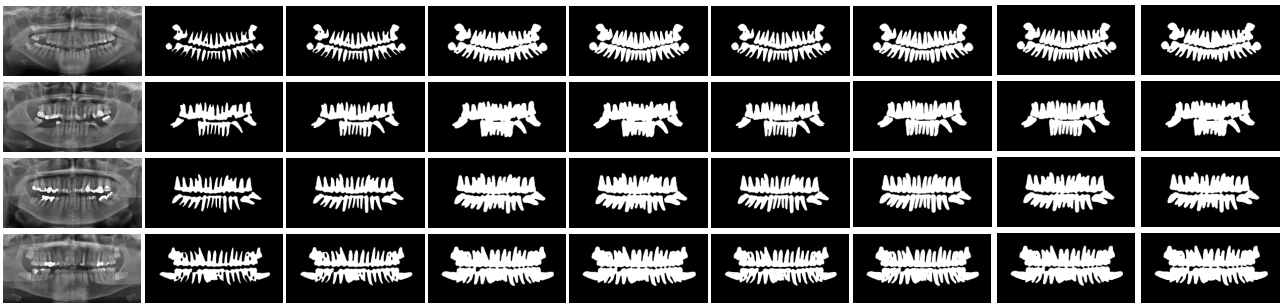

Figure 4. Visual quality comparison on the test data of different models.

4.2. Ablation study

To further validate the effectiveness of CBAM module and deep supervision strategy, we conduct extensive ablation experiments, as detailed in Table 2. Models a, b, c, d denote classical TransUNet, TransUNet+CBAM, TransUNet+CBAM+Concat, TransUNet+Concat+DeepSupervision, TransUNet+CBAM+Concat+DeepSupervision, respectively. The "Concat" means concat the input with the last output feature.

Effectiveness of CBAM: The comparison between models a and b, as well as models d and e, reveals that CBAM contributes to some improvement in the model's capabilities. Due to the ability of CBAM to dynamically adjust the importance of channels and spatial locations in the feature maps generated by CNNs. Through channel attention, it highlights crucial channels, emphasizing informative ones while downplaying less relevant ones. Furthermore, spatial attention allows the model to focus on significant regions within an image. This adaptive recalibration enhances feature representation, making CBAM effective in diverse computer vision tasks.

Table 2. Results of ablation study. Bold indicates the best result.

	Fully supervised			Semi-supervised		
	Dice	IoU	Hausdorff distance	Dice	IoU	Hausdorff distance
a	0.9107	0.9559	0.0315	0.9033	0.9589	0.0317
b	0.9189	0.9588	0.0308	0.9106	0.9601	0.0301
c	0.9206	0.9637	0.0306	0.9135	0.9634	0.0298
d	0.9306	0.9657	0.0300	0.9201	0.9698	0.0286
e	**0.9318**	**0.9691**	**0.0298**	**0.9206**	**0.9723**	**0.0269**

Effectiveness of deep supervision: According to the comparison between models c and e, deep supervision strategy plays an important role in our proposed STS-TransUNet. On the Dice metric, the adoption of the deep supervision strategy shows significant improvement in both full supervision and semi-supervised training. By introducing supervisory signals at multiple layers, deep supervision enables more effective learning of hierarchical features. In turn, this contributes to improved convergence during training and enhances the model's ability to capture intricate patterns in the data.

4.3. Competition results

We participated in MICCAI 2023 Challenges STS-2D Competition with STS-TransUNet and achieved top 3% rankings in both the fully supervised (first round) and semi-supervised (second round) tracks. The detailed results are as follows:

Table 3. Results of competition online test.

	Fully supervised			Semi-supervised		
	Dice	IoU	Hausdorff distance	Dice	IoU	Hausdorff distance
Ours	**0.9334**	**0.9686**	**0.0299**	**0.9113**	**0.9746**	**0.0265**

5. Conclusions

We outline the methodology for both fully and semi-supervised learning with panoramic dental images, covering dataset, partitioning, preprocessing, network architecture, training, comparisons, and evaluation metrics.

We harness a high-quality dataset from various institutions and employ general data preprocessing techniques to ensure the performance and robustness of our model. Furthermore, we seamlessly merge fully supervised and semi-supervised learning, effectively harnessing both labeled and unlabeled data.

We employ a U-shape architecture and introduce a hybrid encoder merging CNN and Transformer strengths, enhancing positional awareness and long-range information fusion. Additionally, CBAM is incorporated to improve spatial and channel information management, contributing to exceptional performance. We train the model in two stages: First, with fully supervised training for a robust baseline, and then transition to semi-supervised training. The semi-supervised approach includes a 'self-training' strategy with pseudo-labels, data augmentation, and iterative model optimization, effectively improving performance with limited labeled data. For evaluation, we compare our model with others in the field. The results unequivocally show its superiority across various metrics, excelling in detail representation, tooth segmentation completeness, and global modeling capabilities. This reaffirms the soundness of our model's design.

Our research has limitations, such as the omission of prior clinical dental knowledge in the model construction. We have focused on the model's architectural priors, inadvertently overlooking the integration of valuable clinical insights. In our future work, we plan to adopt a more inclusive approach, incorporating a broader spectrum of clinical priors to infuse the model with greater real-world clinical relevance and accuracy.

In conclusion, our comprehensive methodology, diverse materials, and rigorous evaluation highlight the outstanding performance of our model in dental panoramic image segmentation. The innovative fusion of CNN and Transformer technologies, along with the implementation of semi-supervised training, establishes it as a front-runner in the field. This study not only provides valuable insights into deep learning applications in medical imaging but also underscores the potential of semi-supervised learning with unlabeled data. In the future, we aim to enhance the practical deployment of our model by integrating clinical information, ensuring that it not only excels in theoretical performance but also demonstrates greater real-world clinical efficacy and relevance.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Students' Innovation and Entrepreneurship Foundation of USTC (No. XY2023S007).

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. P. Sanchez, B. Everett, Y. Salamonson, S. Ajwani, S. Bhole, J. Bishop, et al., Oral health and cardiovascular care: Perceptions of people with cardiovascular disease, *PLoS One*, **12** (2017), e0181189. https://doi.org/10.1371/journal.pone.0181189

2. M. P. Muresan, A. R. Barbura, S. Nedevschi, Teeth detection and dental problem classification in panoramic x-ray images using deep learning and image processing techniques, in *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*, IEEE, (2020), 457–463. https://doi.org/10.1109/ICCP51029.2020.9266244

3. V. Hingst, M. A. Weber, Dental X-ray diagnostics with the orthopantomography–technique and typical imaging results, *Der Radiologe*, **60** (2020), 77–92. https://doi.org/10.1007/s00117-019-00620-1

4. J. C. M. Román, V. R. Fretes, C. G. Adorno, R. G. Silva, J. L. V. Noguera, H. Legal-Ayala, et al., Panoramic dental radiography image enhancement using multiscale mathematical morphology, *Sensors*, **21** (2021), 3110. https://doi.org/10.3390/s21093110

5. R. Izzetti, M. Nisi, G. Aringhieri, L. Crocetti, F. Graziani, C. Nardi, Basic knowledge and new advances in panoramic radiography imaging techniques: A narrative review on what dentists and radiologists should know, *Appl. Sci.*, **11** (2021), 7858. https://doi.org/10.3390/app11177858

6. Y. Zhao, P. Li, C. Gao, Y. Liu, Q. Chen, F. Yang, et al., Tsasnet: Tooth segmentation on dental panoramic X-ray images by Two-Stage Attention Segmentation Network, *Knowledge-Based Syst.*, **206** (2020), 106338. https://doi.org/10.1016/j.knosys.2020.106338

7. A. E. Yüksel, S. Gültekin, E. Simsar, Ş. D. Özdemir, M. Gündoğar, S. B. Tokgöz, et al., Dental enumeration and multiple treatment detection on panoramic X-rays using deep learning, *Sci. Rep.*, **11** (2021), 12342. https://doi.org/10.1038/s41598-021-90386-1

8. R. J. Lee, A. Weissheimer, J. Pham, L. Go, L. M. de Menezes, W. R. Redmond, et al., Three-dimensional monitoring of root movement during orthodontic treatment, *Am. J. Orthod. Dentofacial Orthop.*, **147** (2015), 132–142. https://doi.org/10.1016/j.ajodo.2014.10.010

9. J. Keustermans, D. Vandermeulen, P. Suetens, Integrating statistical shape models into a graph cut framework for tooth segmentation, in *Machine Learning in Medical Imaging*, Springer, (2012), 242–249. https://doi.org/10.1007/978-3-642-35428-1_30

10. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, Springer, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

11. W. Wang, X. Yu, B. Fang, Y. Zhao, Y. Chen, W. Wei, et al., Cross-modality LGE-CMR segmentation using image-to-image translation based data augmentation, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **20** (2023), 2367–2375. https://doi.org/10.1109/tcbb.2022.3140306

12. W. Wang, J. Chen, J. Wang, J. Chen, J. Liu, Z. Gong, Trust-enhanced collaborative filtering for personalized point of interests recommendation, *IEEE Trans. Ind. Inf.*, **16** (2020), 6124–6132. https://doi.org/10.1109/tii.2019.2958696

13. B. G. He, B. Lin, H. P. Li, S. Q. Zhu, Suggested method of utilizing soil arching for optimizing the design of strutted excavations, *Tunnelling Underground Space Technol.*, **143** (2024), 105450. https://doi.org/10.1016/j.tust.2023.105450

14. J. Chen, S. Sun, L. Zhang, B. Yang, W. Wang, Compressed sensing framework for heart sound acquisition in internet of medical things, *IEEE Trans. Ind. Inf.*, **18** (2022), 2000–2009. https://doi.org/10.1109/tii.2021.3088465

15. J. Chen, W. Wang, B. Fang, Y. Liu, K. Yu, V. C. M. Leung, et al., Digital twin empowered wireless healthcare monitoring for smart home, *IEEE J. Sel. Areas Commun.*, **41** (2023), 3662–3676. https://doi.org/10.1109/jsac.2023.3310097

16. Y. Zhang, X. Wu, S. Lu, H. Wang, P. Phillips, S. Wang, Smart detection on abnormal breasts in digital mammography based on contrast-limited adaptive histogram equalization and chaotic adaptive real-coded biogeography-based optimization, *Simulation*, **92** (2016), 873–885. https://doi.org/10.1177/0037549716667834

17. J. H. Lee, S. S. Han, Y. H. Kim, C. Lee, I. Kim, Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs, *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.*, **129** (2020), 635–642. https://doi.org/10.1016/j.oooo.2019.11.007

18. J. Chen, Z. Guo, X. Xu, L. Zhang, Y. Teng, Y. Chen, et al., A robust deep learning framework based on spectrograms for heart sound classification, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **2023** (2023), 1–12. https://doi.org/10.1109/TCBB.2023.3247433

19. S. H. Wang, D. R. Nayak, D. S. Guttery, X. Zhang, Y. D. Zhang, COVID-19 classification by CCSHNet with deep fusion using transfer learning and discriminant correlation analysis, *Inf. Fusion*, **68** (2021), 131–148. https://doi.org/10.1016/j.inffus.2020.11.005

20. H. Chen, X. Huang, Q. Li, J. Wang, B. Fang, J. Chen, Labanet: Lead-assisting backbone attention network for oral multi-pathology segmentation, in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, (2023), 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094785

21. L. Wang, Y. Gao, F. Shi, G. Li, K. C. Chen, Z. Tang, et al., Automated segmentation of dental cbct image with prior-guided sequential random forests, *Med. Phys.*, **43** (2016), 336–346. https://doi.org/10.1118/1.4938267

22. S. Liao, S. Liu, B. Zou, X. Ding, Y. Liang, J. Huang, et al., Automatic tooth segmentation of dental mesh based on harmonic fields, *Biomed Res. Int.*, **2015** (2015). https://doi.org/10.1155/2015/187173

23. R. Girshick, Fast R-CNN, in *Proceedings of the IEEE International Conference on Computer Vision*, (2015), 1440–1448. https://doi.org/10.1109/ICCV.2015.169

24. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 2961–2969. https://doi.org/10.1109/ICCV.2017.322

25. E. Y. Park, H. Cho, S. Kang, S. Jeong, E. Kim, Caries detection with tooth surface segmentation on intraoral photographic images using deep learning, *BMC Oral Health*, **22** (2022), 1–9. https://doi.org/10.1186/s12903-022-02589-1

26. G. Zhu, Z. Piao, S. C. Kim, Tooth detection and segmentation with mask R-CNN, in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, IEEE, (2020), 070–072. https://doi.org/10.1109/ICAIIC48513.2020.9065216

27. Q. Chen, Y. Zhao, Y. Liu, Y. Sun, C. Yang, P. Li, et al., Mslpnet: Multi-scale location perception network for dental panoramic X-ray image segmentation, *Neural Comput. Appl.*, **33** (2021), 10277–10291. https://doi.org/10.1007/s00521-021-05790-5

28. P. Li, Y. Liu, Z. Cui, F. Yang, Y. Zhao, C. Lian, et al., Semantic graph attention with explicit anatomical association modeling for tooth segmentation from CBCT images, *IEEE Trans. Med. Imaging*, **41** (2022), 3116–3127. https://doi.org/10.1109/tmi.2022.3179128
29. E. Shaheen, A. Leite, K. A. Alqahtani, A. Smolders, A. Van Gerven, H. Willems, et al., A novel deep learning system for multi-class tooth segmentation and classification on Cone Beam Computed Tomography. A validation study, *J. Dent.*, **115** (2021), 103865. https://doi.org/10.1016/j.jdent.2021.103865
30. M. Ezhov, A. Zakirov, M. Gusarev, Coarse-to-fine volumetric segmentation of teeth in cone-beam CT, in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, (2019), 52–56. https://doi.org/10.1109/ISBI.2019.8759310
31. A. Alsheghri, F. Ghadiri, Y. Zhang, O. Lessard, J. Keren, F. Cheriet, et al., Semi-supervised segmentation of tooth from 3D scanned dental arches, in *Medical Imaging 2022: Image Processing*, (2022), 766–771. https://doi.org/10.1117/12.2612655
32. X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, et al., Self-supervised learning: Generative or contrastive, *IEEE Trans. Knowl. Data Eng.*, **35** (2021), 857–876. https://doi.org/10.1109/tkde.2021.3090866
33. Q. Li, X. Huang, Z. Wan, L. Hu, S. Wu, J. Zhang, et al., Data-efficient masked video modeling for self-supervised action recognition, in *Proceedings of the 31st ACM International Conference on Multimedia*, (2023), 2723–2733. https://doi.org/10.1145/3581783.3612496
34. H. Lim, S. Jung, S. Kim, Y. Cho, I. Song, Deep semi-supervised learning for automatic segmentation of inferior alveolar nerve using a convolutional neural network, *BMC Oral Health*, **21** (2021), 1–9. https://doi.org/10.2196/preprints.32088
35. F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, K. H. Maier-Hein, nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation, *Nat. Methods*, **18** (2021), 203–211. https://doi.org/10.1038/s41592-020-01008-z
36. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint*, (2020), arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929
37. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in *International Conference on Machine Learning*, PMLR, (2021), 10347–10357.
38. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in *European Conference on Computer Vision*, Springer, (2020), 213–229. https://doi.org/10.1007/978-3-030-58452-8_13
39. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: Deformable transformers for end-to-end object detection, *arXiv preprint*, (2020), arXiv:2010.04159. https://doi.org/10.48550/arXiv.2010.04159
40. Q. Li, X. Huang, B. Fang, H. Chen, S. Ding, X. Liu, Embracing large natural data: Enhancing medical image analysis via cross-domain fine-tuning, *IEEE J. Biomed. Health. Inf.*, **2023** (2023), 1–10. https://doi.org/10.1109/JBHI.2023.3343518

41. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, et al., Transunet: Transformers make strong encoders for medical image segmentation, *arXiv preprint*, arXiv:2102.04306. https://doi.org/10.48550/arXiv.2102.04306

42. A. Srinivas, T. Lin, N. Parmar, J. Shlens, P. Abbeel, A. Vaswani, Bottleneck transformers for visual recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 16519–16529. https://doi.org/10.1109/CVPR46437.2021.01625

43. Y. Li, S. Wang, J. Wang, G. Zeng, W. Liu, Q. Zhang, et al., GT U-Net: A U-Net like group transformer network for tooth root segmentation, in *Machine Learning in Medical Imaging*, Springer, (2021), 386–395. https://doi.org/10.1007/978-3-030-87589-3_40

44. W. Lin, Z. Wu, J. Chen, J. Huang, L. Jin, Scale-aware modulation meet transformer, *arXiv preprint*, (2023), arXiv:2307.08579. https://doi.org/10.48550/arXiv.2307.08579

45. S. Woo, J. Park, J. Lee, I. Kweon, CBAM: Convolutional block attention module, in *Computer Vision–ECCV 2018*, Springer, (2018), 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

46. Y. Zhang, F. Ye, L. Chen, F. Xu, X. Chen, H. Wu, et al., Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection, *Sci. Data*, **10** (2023), 380. https://doi.org/10.1038/s41597-023-02237-5

47. K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, F. Nie, A semisupervised recurrent convolutional attention model for human activity recognition, *IEEE Trans. Neural Networks Learn. Syst.*, **31** (2019), 1747–1756. https://doi.org/10.1109/tnnls.2019.2927224

48. G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, et al., Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, *Sci. Rep.*, **6** (2016), 26286. https://doi.org/10.1038/srep26286

49. J. Chen, L. Chen, Y. Zhou, Cryptanalysis of a DNA-based image encryption scheme, *Inf. Sci.*, **520** (2020), 130–141. https://doi.org/10.1016/j.ins.2020.02.024

50. D. Yuan, X. Chang, P. Y. Huang, Q. Liu, Z. He, Self-supervised deep correlation tracking, *IEEE Trans. Image Process.*, **30** (2020), 976–985. https://doi.org/10.1109/tip.2020.3037518

51. Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, Z. Liang, Apple detection during different growth stages in orchards using the improved YOLO-V3 model, *Comput. Electron. Agric.*, **157** (2019), 417–426. https://doi.org/10.1016/j.compag.2019.01.012

52. D. Yuan, X. Chang, Q. Liu, Y. Yang, D. Wang, M. Shu, et al., Active learning for deep visual tracking, *IEEE Trans. Neural Networks Learn. Syst.*, **2023** (2023), 1–13. https://doi.org/10.1109/TNNLS.2023.3266837

53. Y. Zhang, L. Deng, H. Zhu, W. Wang, Z. Ren, Q. Zhou, et al., Deep learning in food category recognition, *Inf. Fusion*, **98** (2023), 101859. https://doi.org/10.1016/j.inffus.2023.101859

54. D. Cremers, M. Rousson, R. Deriche, A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape, *Int. J. Comput. Vision*, **72** (2007), 195–215. https://doi.org/10.1007/s11263-006-8711-1

55. X. Shu, Y. Yang, J. Liu, X. Chang, B. Wu, Alvls: Adaptive local variances-based levelset framework for medical images segmentation, *Pattern Recognit.*, **136** (2023), 109257. https://doi.org/10.1016/j.patcog.2022.109257

56. K. Ding, L. Xiao, G. Weng, Active contours driven by region-scalable fitting and optimized laplacian of gaussian energy for image segmentation, *Signal Process.*, **134** (2017), 224–233. https://doi.org/10.1016/j.sigpro.2016.12.021

57. G. Jader, J. Fontineli, M. Ruiz, K. Abdalla, M. Pithon, L. Oliveira, Deep instance segmentation of teeth in panoramic X-ray images, in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, IEEE, (2018), 400–407. https://doi.org/10.1109/SIBGRAPI.2018.00058

58. T. L. Koch, M. Perslev, C. Igel, S. S. Brandt, Accurate segmentation of dental panoramic radiographs with U-Nets, in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, (2019), 15–19. https://doi.org/10.1109/ISBI.2019.8759563

59. Z. Cui, C. Li, N. Chen, G. Wei, R. Chen, Y. Zhou, et al., Tsegnet: An efficient and accurate tooth segmentation network on 3D dental model, *Med. Image Anal.*, **69** (2021), 101949. https://doi.org/10.1016/j.media.2020.101949

60. X. Wang, S. Gao, K. Jiang, H. Zhang, L. Wang, F. Chen, et al., Multi-level uncertainty aware learning for semi-supervised dental panoramic caries segmentation, *Neurocomputing*, **540** (2023), 126208. https://doi.org/10.1016/j.neucom.2023.03.069

61. A. Qayyum, A. Tahir, M. A. Butt, A. Luke, H. T. Abbas, J. Qadir, et al., Dental caries detection using a semi-supervised learning approach, *Sci. Rep.*, **13** (2023), 749. https://doi.org/10.1038/s41598-023-27808-9

62. Y. Zhang, H. Liu, Q. Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021*, Springer, (2021), 14–24. https://doi.org/10.1007/978-3-030-87193-2_2

63. Y. Wang, T. Wang, H. Li, H. Wang, ACF-TransUNet: Attention-based coarse-fine transformer U-Net for automatic liver tumor segmentation in CT images, in *2023 4th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, IEEE, (2023), 84–88. https://doi.org/10.1109/ICBASE59196.2023.10303169

64. B. Chen, Y. Liu, Z. Zhang, G. Lu, A. W. K. Kong, TransAttUnet: Multi-level attention-guided U-Net with transformer for medical image segmentation, *IEEE Trans. Emerging Top. Comput. Intell.*, **2023** (2023), 1–14. https://doi.org/10.1109/TETCI.2023.3309626

65. A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, et al., UNETR: Transformers for 3D medical image segmentation, in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (2022), 1748–1758. https://doi.org/10.1109/WACV51458.2022.00181

66. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, et al., Swin-Unet: Unet-like pure transformer for medical image segmentation, in *European Conference on Computer Vision*, Springer, (2022), 205–218. https://doi.org/10.1007/978-3-031-25066-8_9

67. S. Li, C. Li, Y. Du, L. Ye, Y. Fang, C. Wang, et al., Transformer-based tooth segmentation, identification and pulp calcification recognition in CBCT, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, (2023), 706–714. https://doi.org/10.1007/978-3-031-43904-9_68

68. M. Kanwal, M. M. Ur Rehman, M. U. Farooq, D. K. Chae, Mask-transformer-based networks for teeth segmentation in panoramic radiographs, *Bioengineering*, **10** (2023), 843. https://doi.org/10.3390/bioengineering10070843

69. W. Chen, X. Du, F. Yang, L. Beyer, X. Zhai, T. Y. Lin, et al., A simple single-scale vision transformer for object detection and instance segmentation, in *European Conference on Computer Vision*, Springer, (2022), 711–727. https://doi.org/10.1007/978-3-031-20080-9_41

70. M. R. Amini, V. Feofanov, L. Pauletto, E. Devijver, Y. Maximov, Self-training: A survey, *arXiv preprint*, (2023), arXix: 2202.12040. https://api.semanticscholar.org/CorpusID:247084374

71. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: A nested U-Net architecture for medical image segmentation, in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, (2018), 3–11. https://doi.org/10.1007/978-3-030-00889-5_1

72. H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, et al., Unet 3+: A full-scale connected unet for medical image segmentation, in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, IEEE, (2020), 1055–1059. https://doi.org/10.1109/ICASSP40776.2020.9053405

73. Q. Zuo, S. Chen, Z. Wang, R2AU-Net: Attention recurrent residual convolutional neural network for multimodal medical image segmentation, *Secur. Commun. Netw.*, **2021** (2021), 1–10. https://doi.org/10.1155/2021/6625688

74. C. Sheng, L. Wang, Z. Huang, T. Wang, Y. Guo, W. Hou, et al., Transformer-based deep learning network for tooth segmentation on panoramic radiographs, *J. Syst. Sci. Complexity*, **36** (2023), 257–272. https://doi.org/10.1007/s11424-022-2057-9

75. R. Azad, R. Arimond, E. K. Aghdam, A. Kazerouni, D. Merhof, DAE-former: Dual attention-guided efficient transformer for medical image segmentation, in *Predictive Intelligence in Medicine*, Springer, (2023), 83–95. https://doi.org/10.1007/978-3-031-46005-0_8

76. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, Segformer: Simple and efficient design for semantic segmentation with transformers, *Adv. Neural Inf. Process. Syst.*, **34** (2021), 12077–12090.

AIMS Press