Mathematical Biosciences and Engineering

MBE, 20(5): 8320–8336.
DOI: 10.3934/mbe.2023364
Received: 03 February 2023
Revised: 18 February 2023
Accepted: 26 February 2023
Published: 01 March 2023

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

Research article

CoT-UNet++: A medical image segmentation method based on contextual transformer and dense connection

Yijun Yin[1], Wenzheng Xu[1], Lei Chen[1,*] and Hao Wu[2,*]

[1] School of Information Science and Engineering, Shandong University, Qingdao 266200, China
[2] Department of Stomatology, the First Medical Centre, Chinese PLA General Hospital, Beijing 100853, China

***Correspondence:** Email: lei.chen@sdu.edu.cn.

Abstract: Accurate depiction of individual teeth from CBCT images is a critical step in the diagnosis of oral diseases, and the traditional methods are very tedious and laborious, so automatic segmentation of individual teeth in CBCT images is important to assist physicians in diagnosis and treatment. TransUNet has achieved success in medical image segmentation tasks, which combines the advantages of Transformer and CNN. However, the skip connection taken by TransUNet leads to unnecessary restrictive fusion and also ignores the rich context between adjacent keys. To solve these problems, this paper proposes a context-transformed TransUNet++ (CoT-UNet++) architecture, which consists of a hybrid encoder, a dense connection, and a decoder. To be specific, a hybrid encoder is first used to obtain the contextual information between adjacent keys by CoTNet and the global context encoded by Transformer. Then the decoder upsamples the encoded features by cascading upsamplers to recover the original resolution. Finally, the multi-scale fusion between the encoded and decoded features at different levels is performed by dense concatenation to obtain more accurate location information. In addition, we employ a weighted loss function consisting of focal, dice, and cross-entropy to reduce the training error and achieve pixel-level optimization. Experimental results demonstrate that the proposed CoT-UNet++ method outperforms the baseline models and can obtain better performance in tooth segmentation.

Keywords: medical image segmentation; tooth segmentation; contextual transformer; dense connection; weighted loss function

1. Introduction

With the progress of society and the improvement of living standards, people are increasingly aware of oral health care. More people actively carry out dental treatment, such as orthodontics and dental implants, to ensure the normal function of teeth and improve facial appearance [1,2]. Orthodontic treatment not only beautifies the alignment of teeth but also affects the related soft tissues, and improves the aesthetic appearance of the patient's face. In addition, severe endodontic and periapical diseases need to be treated with the help of dental implant technology. During dental treatment, 3D cone-beam computed tomography (CBCT) images are commonly used to assist in diagnosis. It can fulfill the dentist's need to observe the root alignment of the patient's teeth and can provide comprehensive 3D information on the complete tooth. Traditional dental treatment relies on the subjective experience of the doctor to analyze the spatial alignment and morphology of the patient's crowns, root canals, and alveolar bone through CBCT images. The process of manually measuring teeth in CBCT images can consume a lot of time and energy for the dentist. Therefore, a tooth segmentation algorithm is expected to be designed and used in the clinic. The input of a patient's CBCT image can automatically segment the tooth part precisely.

Nowadays, artificial intelligence and deep learning have gained breakthroughs and become the focus of public attention. Computer vision is the direction with the longest research history and the most accumulated technology in artificial intelligence, and it has shown strong applications [3–12] in various fields such as image classification, image segmentation, super-resolution, and human face, and many problems of medical image segmentation have been solved. Therefore, we applied computer vision to dental processing to achieve high-precision segmentation of CBCT dental images and to exploit its value in the clinic.

Among the segmentation methods for medical images, U-Net and its variants consisting of an encoder-decoder with skip connections have shown good segmentation performance [13–15]. The skip connection is able to combine coarse-grained and fine-grained feature maps for better segmentation of images. Based on this approach, great success has been achieved in a wide range of medical applications, such as cardiac segmentation by magnetic resonance (MR) [16], organ segmentation by computed tomography (CT) [17–19], and polyp segmentation [15]. Based on this structure, various methods such as UNet + + [20], UNet 3 + [21], DenseUNet [17], and KiU-Net [22] were subsequently proposed and have also been successful in the segmentation of various medical imaging datasets. UNet++ redesigns the fusion scheme of different scale features based on UNet to achieve highly flexible feature fusion. Although the CNN-based methods have shown excellent performance, UNet has difficulty in handling long-range and global semantic information, especially when there is low contrast between the organ and the environment, due to the intrinsic local nature of the convolution operation. Therefore, researchers have built a global context model by using a self-attention mechanism [23,24], but still cannot fundamentally solve the above problem.

Many traditional tooth segmentation algorithms have been proposed, which reflects the importance of these applications. Previous approaches used region growth [25], level sets boosted variants [26,27] and statistical shape models [28,29], but these methods suffered many failures. A new region candidate network [30] can effectively remove duplicate candidates and speed up the training, but also ends up with some erroneous segmentation.

In recent years, the Transformer architecture, which has achieved great success in the field of Natural Language Processing (NLP), has been gradually applied to the field of Computer Vision

(CV). Designed for sequence-to-sequence prediction and relying entirely on self-attentive mechanisms, it has shown power in modeling global context, while conversely lacking attention to low-resolution information in images. TransUNet [31], as a classical hybrid network, combines the advantages of UNet and Transformer, not only has adopted global information, but also can extract low-level information to compensate for fine details, and has shown strong performance in medical image segmentation. After this, new architectures combining CNN and Transformer modules have been continuously developed [32–38]. However, these traditional self-attention methods only use query-key pairs to compute the attention matrix without fully considering the rich context between keys, and Li et al. [39] proposed the CoTNet to address this problem. The CoTNet is obtained after replacing the 3×3 convolution in ResNet with the CoT block, which achieves good performance in different tasks such as target detection and instance segmentation.

Inspired by the CoT block of the CoTNet [39] and UNet++ [20], we propose CoT-UNet++, which fully uses contextual information between keys. Moreover, we adopt a densely connected fusion scheme. This learns not only local contextual information, but also adopts feature maps at all scales of the encoder, and obtains accurate localization information. In summary, the contributions of this paper are as follows:

- A CoT block is utilized in the encoder section to make full use of the contextual information between keys, thus enhancing visual representations.
- A dense connection is introduced to leverage multi-scale features that combine low-level detail and high-level semantics in full-scale feature maps.
- A weighted loss function consisting of cross-entropy loss, Dice loss, and focus loss is used to reduce the training error of medical image segmentation and achieve pixel-level optimization.

2. The proposed CoT-UNet++ model

2.1. General architecture of CoT-UNet++

The overall structure of CoT-UNet++ is shown in Figure 1, which mainly consists of a hybrid encoder, dense connection, and a decoder.

As can be seen in Figure 1(b), CoT-UNet++ is a tightly mixed CNN and Transformer model that first feeds the input image into the hybrid encoder and uses CoTNet-50 to extract features with different resolutions. Part of the structure in CoTNet-50 is shown in Figure 1(a), and the specific structure is described in Section 3.2.2. The CoT block makes full use of the contextual information between keys and can facilitate self-attention learning. In the final stage of CoTNet-50, the feature map is mapped into the form of a sequence and the embedded position information is fed into the Transformer encoder for multi-head self-attention learning, which achieves long-range dependency to obtain global encoded features. The transformed encoded features are then reshaped into feature maps for decoding, and a cascaded upsampler is used to decode the low-resolution features, with each upsampling block consisting of 3×3 convolution and ReLU layers.

To achieve flexible feature fusion, a dense connection is applied to fuse both the feature maps of the encoder output and intermediate nodes with the ones obtained by decoding, which can combine multiple low-level detailed feature maps from the encoding side with the feature maps containing high-level semantic information from the decoding side. The intermediate nodes use a multi-scale combination strategy of 1×1 and 3×3 convolution to capture more different local features.

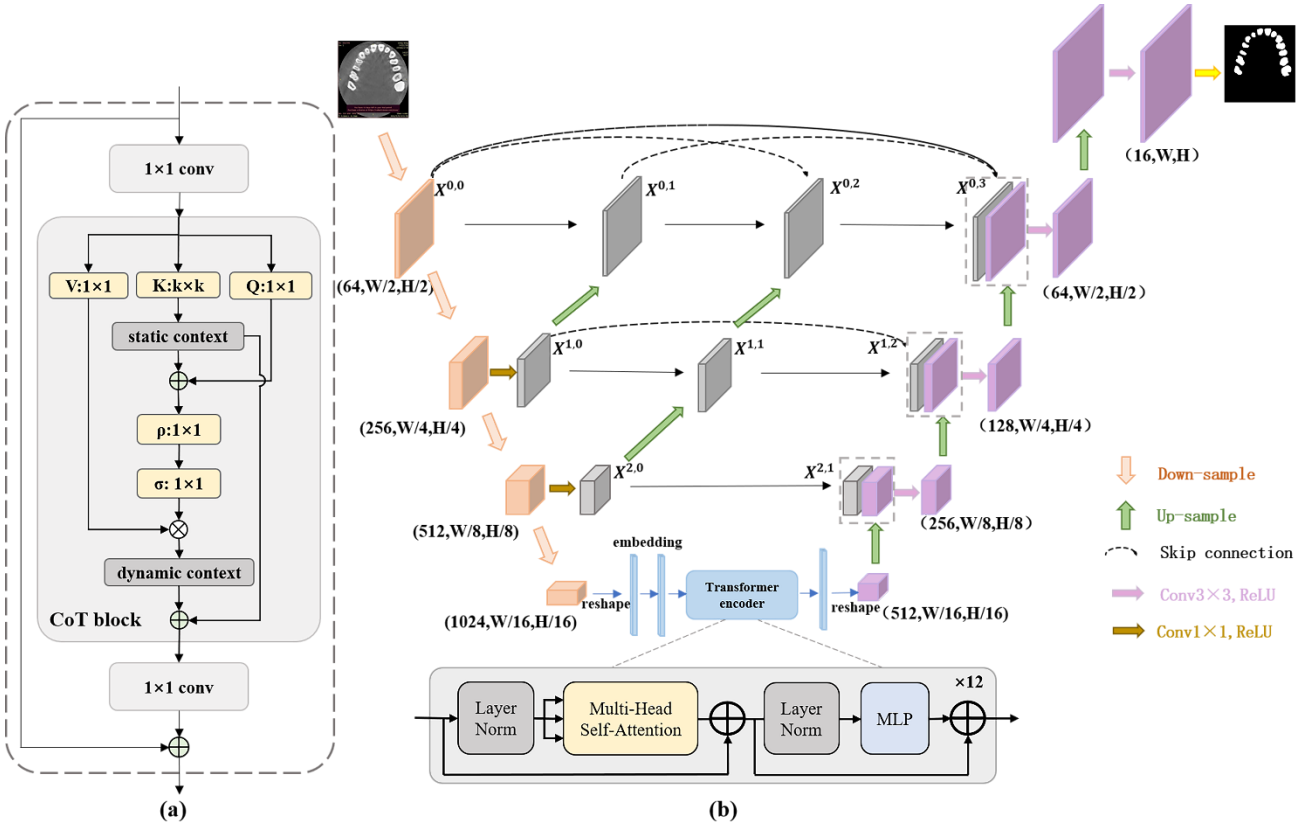

Figure 1. The overall structure of CoT-UNet++. (a) The specific structure of the CoT block. (b) Overall architecture of CoT-Unet.

2.2. Hybrid encoder

2.2.1. Multi-head self-attention

The traditional self-attention mechanism first transforms the input 2D feature mapping graph X into three different matrices queries, keys and values, that is $Q = XW^Q$, $K = XW^K$, $V = XW^V$, which are computed from the input vectors with their weight matrices W^Q, W^K, W^V. As shown in Eq (1), the similarity matrix QK^T is first obtained by multiplying the transpose of Q with K. The weight matrix is obtained by normalizing it with the softmax function. Finally, the attention matrix of the input vector $Attention$ is obtained by multiplying the weight matrix with V.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \times V \tag{1}$$

where d_k denotes the dimensionality of the Q or K matrix.

The multi-head self-attention mechanism is an important part of the Transformer, which is a combination of m self-attention modules. Q, K, V correspond to m weight matrices, where W_i^Q, W_i^K, W_i^V denote the weight matrix of the ith self-attention respectively. They are multiplied with the input vector X^i to obtain the projection matrices $Q_i = QW_i^Q$, $K_i = KW_i^K$ and $V_i = VW_i^V$ on different spaces, and the attention matrix of each head is calculated as in Eq (2). Then all the output matrices are stitched together and multiplied with the learnable linear transformation matrix W^O to obtain the final multi-head self-attention output matrix MSA, shown as Eq (3).

$$head_i = Attention(Q_i, K_i, V_i) \tag{2}$$

$$MSA(Q, K, V) = Concat(head_1, \ldots head_m)W^O \tag{3}$$

2.2.2. CoT block

In the traditional self-attention mechanism, all query-key pairs are learned independently without using rich contextual information. For this purpose, the CoT block is constructed to combine the contextual information of adjacent keys and self-attention learning into a single structure to enhance the representational power of the learned feature graph. The specific structure of the CoT block is shown in Figure 1(a). It is assumed that in an input 2D feature mapping graph X, its queries, keys, and values are Q, K, V, respectively. The traditional self-attention mechanism directly multiplies the transpose of the query matrix and the key matrix by QK^T to derive the similarity matrix between them. However, in the CoT block, to achieve the contextual representation of each key, the adjacent keys are first convolved with k × k to obtain K^1, a representation of static contextual information. As shown in Eq (4), K^1 of the learned static context information is spliced with Q, and then two consecutive 1 × 1 convolutions (W_1 with ReLU activation function and W_2 without activation function) are performed to obtain the attention matrix $Attention$.

$$Attention = (K^1 \oplus Q)W_1W_2 \tag{4}$$

For each head, the local attention matrix for each spatial location is learned from Q and the integrated contextual information of K instead of isolated query-key pairs, which enhances self-attention learning. The obtained attention matrix $Attention$ is then multiplied with V to obtain the dynamic contextual representation K^2, shown as Eq (5).

$$K^2 = V \otimes Attention \tag{5}$$

Finally, the static and dynamic contextual information is combined as the final output of the CoT block, which is shown in Eq (6).

$$Y = K^1 + K^2 \tag{6}$$

Table 1. The detailed structures for ResNet-50 and CoTNet-50.

stage	ResNet-50	CoTNet-50	output
res1	7 × 7 conv, 64, stride 2	7 × 7 conv, 64, stride 2	112 × 112
	3 × 3 max pool, stride 2	3 × 3 max pool, stride 2	
res2	$\begin{bmatrix} 1\times1,64 \\ 3\times3,64 \\ 1\times1,256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1,64 \\ \mathbf{CoT,64} \\ 1\times1,256 \end{bmatrix}\times3$	56 × 56
res3	$\begin{bmatrix} 1\times1,128 \\ 3\times3,128 \\ 1\times1,512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1,128 \\ \mathbf{CoT,128} \\ 1\times1,512 \end{bmatrix}\times4$	28 × 28
res4	$\begin{bmatrix} 1\times1,256 \\ 3\times3,256 \\ 1\times1,1024 \end{bmatrix}\times9$	$\begin{bmatrix} 1\times1,256 \\ \mathbf{CoT,256} \\ 1\times1,1024 \end{bmatrix}\times9$	14 × 14

In summary, we obtained the above two kinds of contextual information by capturing information between adjacent keys. And visual representation learning is promoted by static contextual information obtained through 3 × 3 convolution and dynamic contextual information

based on static context for self-attention learning.

We partially modified ResNet-50 by combining the res4 and res5 layers into one layer and replacing the 3 × 3 convolution of them with a CoT block to obtain CoTNet-50. The structure of CoTNet-50 is shown in Table 1.

2.2.3. Vision Transformer block

For an image $X \in \Phi^{\mathrm{H\times W\times}C}$, $H \times W$ is the resolution of the image and C is the number of channels. The standard Transformer receives a 1D sequence as input. So the image is first reconstructed as a 2D patch $\left\{X_P^i \in \Phi^{P^2C} \mid i = 1, \ldots, L\right\}$. $P \times P$ is the resolution of each image patch, $L = \frac{HW}{P^2}$ is the number of patches, which also serves as the input sequence length for the Transformer. Then the vectorized X_P^i is mapped into the N-dimensional space by linear projection. The output of this projection is patch embedding. Finally, the position embedding is added to the patch embedding to preserve the position information. The resulting sequence of embedding vectors is used as the input to the Transformer encoder. The above process can be expressed by Eq (7).

$$Y_0 = \left[X_P^1 I; X_P^2 I; \ldots X_P^L I\right] + I_p \tag{7}$$

where $I \in \Phi^{(P^2C)\times N}$ is the embedding projection of patch, $I_p \in \Phi^{L\times N}$ is the position embedding.

The Transformer layer is composed of an L-layer Multi-head Self-Attention (MSA) and a Multilayer Perceptron (MLP). The multilayer perceptron mainly consists of a linear combination of two fully connected layers and a linear activation layer ReLU. The output of the Lth layer can be expressed by Eqs (8) and (9).

$$Y_l' = MSA\left(LN\left(Y_{l-1}\right)\right) + Y_{l-1} \tag{8}$$

$$Y_l = MLP\left(LN\left(Y_l'\right)\right) + Y_l' \tag{9}$$

where $LN(\cdot)$ denotes the normalization operator, Y_l is the encoded image representation, and Y_{l-1} is the output of a Transformer layer on Y_l.

2.3. Dense connection

Let $x^{i,j}$ denote the output of node $X^{i,j}$, where i is the index of the downsampling layer along the encoder and j is the number of convolutional layers of the intermediate nodes connected along the skip. $x^{i,j}$ can be represented as follows:

$$x^{i,j} = \begin{cases} Conv\left(Down\left(x^{i-1,j}\right)\right) & j = 0 \\ Conv\left(\left[\left[x^{i,k}\right]_{k=0}^{j-1}, Up\left(x^{i+1,j-1}\right)\right]\right) & j > 0 \end{cases} \tag{10}$$

where the function $Conv(\cdot)$ is a convolution operation, each convolution operation is followed by an activation function, $Up(\cdot)$ and $Down(\cdot)$ denote the up-sampling and down-sampling layers, and $[\cdot]$ denotes the connection layer. As shown in Figure 1(b), for $j = 0$, $X^{i+1,0}$ only receives one input and is the previous layer node $X^{i,0}$ of the encoder. When $j = 1$, the $X^{i+1,1}$ node receives two inputs $X^{i,0}$, $X^{i+1,0}$, which are two consecutive sub-networks from the encoder. $j > 1$, the $X^{i+1,j}$ node receives $j + 1$ inputs, where the j inputs $X^{i+1,0}$,…,$X^{i+1,j-1}$ are all the outputs of the first j nodes of the same layer of the skip connection, and the last $j + 1$th input is the upsampled output of the skip connection from the lower layer. When $j > 1$, the $X^{i+1,j}$ node receives $j + 1$ inputs, where the j inputs $X^{i+1,0}$,…,$X^{i+1,j-1}$ are all the outputs of the first j nodes of the same layer of the skip connection, and the last $j + 1$th input is the upsampled output of the skip connection from the lower layer. We use a convolutional block to complete the skip connection between each of the above nodes. This enables all the previous feature maps can be accumulated and reach the current node, which includes not only the final aggregated feature maps, but also the maps of the intermediate nodes and the original feature maps of the same scale from the encoder. In this way, the deep fusion at different levels of features is completed, and the unnecessary restriction behavior of skip connections is further relaxed.

2.4. Weighted loss function

We use a weighted loss function in the model. $\mathcal{L}_{ce}$ is the cross-entropy loss, and can be used to evaluate the loss incurred when classifying pixel points during the segmentation of image data, which can be defined as follows:

$$\mathcal{L}_{ce} = -\frac{1}{n}\left[\sum_i y_i log\widehat{y_i} + (1 - y_i)\log(1 - \widehat{y_i})\right] \tag{11}$$

where y_i denotes the label of sample i, the positive class is 1 and the negative class is 0. $\widehat{y_i}$ denotes the probability that sample i is predicted to be positive class after training. The base of log is e and $\mathcal{L}_{dice}$ is a loss of Dice, which can be used to evaluate the similarity between the predicted segmented image and the real segmented image (label) whose value ranges from [0,1] and is calculated as follows:

$$\mathcal{L}_{dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{12}$$

where $|X \cap Y|$ denotes the intersection of the real picture and the predicted picture , and $|X|$ and $|Y|$ denote the number of the corresponding elements.

The focal loss function [40] $\mathcal{L}_{foc}$ focuses the training on hard negatives samples and is also able to alleviate the problem of unbalanced data samples, and it is calculated as follows:

$$\mathcal{L}_{foc} = -\frac{1}{n}\left[\sum_i (\alpha\left(1 - \widehat{y_i}\right)^{\gamma} y_i log\widehat{y_i} + (1 - \alpha)\widehat{y_i}^{\gamma}\left(1 - y_i\right)\log\left(1 - \widehat{y_i}\right))\right] \tag{13}$$

where α is the weighting factor and takes values in the range [0, 1]. For positive samples, the weight is α, and for negative samples, it is $(1 - \alpha)$. γ is a parameter that takes values in [0,5].

When $\widehat{y}_i$ tends to 1, it indicates that the sample is easily distinguishable and the modulation factor $\left(1-\widehat{y}_i\right)^{\gamma}$ tends to 0, which indicates a smaller contribution to the loss, thus reducing the loss contribution of the easily distinguishable sample. When $\widehat{y}_i$ is small, it may be misclassified into positive samples. At this time the modulation factor $\left(1-\widehat{y}_i\right)^{\gamma}$ converges to 1, which does not have much effect on the loss. By reducing the loss contribution of the easy-to-score samples, the loss obtained for simple samples becomes smaller, while the loss for samples with small prediction probability becomes large, thus enhancing the focus on difficult cases.

We take a weighted combination of cross-entropy loss, Dice loss and focus loss, which can effectively segment the boundaries and overall contours of teeth in medical images. The weighted loss is defined as follows:

$$\mathcal{L} = \mathrm{A}\mathcal{L}_{ce} + B\mathcal{L}_{dice} + C\mathcal{L}_{foc} \tag{14}$$

where A, B and C are the weight parameters of the three loss functions respectively.

3. Experiments

3.1. Dataset and pre-processing

There is no publicly available dataset of CBCT dental images for research, so we first constructed a dataset of CBCT dental images. We collected a large number of CBCT images from dental hospitals, and all dental CBCT images were obtained from patients under routine clinical care. Most of these patients need dental treatment such as dental implants, orthodontics, and restorations. To determine the true labels of the teeth for model training and evaluation, we perform point-by-point labeling of the tooth parts in the CBCT images. The labeling process is fully manually outlined and checked by an experienced physician to ensure the accuracy of the labeling. In total, there are 20 groups of 300 dental CBCT images in the dental dataset were constructed for tooth segmentation experiments.

We normalized the data to a uniform size before training, and then the data is enhanced by cropping, rotating, and mirroring. We selected three sets of different sizes of crop sizes, followed by different angular rotations, and also used up-down mirroring and left-right mirroring, where each transformation was randomly applied to the original image to complete the enhancement of the dataset.

3.2. Implementation and evaluation

3.2.1. Implementation

We performed experiments on the PyTorch platform. The hybrid encoder module used ResNet-50 [41] as the baseline model, where the 3 × 3 convolution was replaced with a CoT block, called CoTNet-50 for feature extraction. And we adopted ViT [34] with 12 Transformer layers and a multi-head self-attention mechanism with 12 heads. We combined CoTNet-50 and ViT, denoted as C50-ViT as a hybrid encoder. The input resolution size of the image was 224 × 224 and the patch was 16. The model was trained by the SGD optimizer with a learning-rate 0.001, momentum 0.9, weight decay 1e-4, epochs 200 and batch-size 4. The parameters α, γ were set to 0.25 and 0.6. And

the loss function weights A, B, and C were 0.4, 0.4 and 0.2 respectively. All experiments were conducted using the NVIDIA GeForce RTX 3060 GPU.

3.2.2. Evaluation metrics

Two commonly-used evaluation metrics in image segmentation are Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (95HD).

Dice Similarity Coefficient (DSC). In medical image segmentation, the Dice Similarity Coefficient is used to calculate the similarity between the model segmentation result and the real label, and the range of the value is [0, 1]. The closer DSC value is to 1, the closer the segmentation result is to the ground truth, which indicates better segmentation performance. The calculation of DSC is as follows:

$$Dice(X,Y) = \frac{2|X \cap Y|}{|X|+|Y|} \tag{15}$$

where X denotes the set of true label pixels and Y denotes the output segmentation result of the model.

Hausdorff distance (HD). Hausdorff distance is used to calculate the distance between any point set $X = \{x1, x2, \ldots xn\}$ in space. Another point set $= \{y1, y2, \ldots . xn\}$, which can be used in image segmentation tasks to measure the maximum mismatch between two contours or shapes. 95HD represents the 95% quantile of the maximum distance, which is slightly more stable for small outliers compared to the Hausdorff distance and is commonly used for biomedical segmentation challenges. It is calculated as follows:

$$H(X,Y) = \max\{\max\ min\|x-y\|, \max\ min\|y-x\|\} \tag{16}$$

The Hausdorff distance represents the maximum value of the difference between the point set X and the point set Y. Therefore, a smaller value represents a better match between the two contours or shapes, and a better segmentation of the model. We also used the following additional evaluation metrics to assess the segmentation results of the model.

Pixel Accuracy (PA). PA is the simplest metric and is the ratio of correctly marked pixels to the total pixels. The calculated formula is as follows:

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=o}^{k} p_{ij}} \tag{17}$$

Mean Pixel Accuracy (MPA). MPA is a simple enhancement of PA, which calculates the proportion of pixels within each class that are correctly classified. The calculation formula is as follows:

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{18}$$

Mean Intersection over Union (mIoU). This metric provides a balanced evaluation of the segmentation results for all categories. The segmentation accuracy is measured by calculating the ratio of intersection and union between the segmentation results and the true labels. The calculation

formula is as follows:

$$IoU = \frac{TP}{FP+TP+FN} \tag{19}$$

Average System Surface Distance (ASD). It calculates the average of the surface distances between the segmentation result and all points in the real label, which is one of the evaluation criteria for the medical image segmentation competition CHAOS. Let the point set $A = \{a1, a2, \dots an\}$ be the surface point set of the model segmentation result, and the point set $B = \{b1, b2, \dots bn\}$ is the surface point set of the real label. The calculation of ASD can be expressed as Formula (20). The unit of ASD is millimeter, and its smaller value means the better segmentation result of the model.

$$ASD = \frac{1}{S(A)+S(B)}\left(\sum_{s_A \in S(A)} d(s_A, S(B)) + \sum_{s_B \in S(B)} d(s_B, S(A)\right) \tag{20}$$

3.3. Performance comparison on teeth segmentation

3.3.1. The selection of loss function weights

We combine three different loss functions to obtain our weighted loss function. The optimal weights of A, B and C are derived from multiple sets of experimental data, as shown in Table 3. We took four different sets of weights, and the experimental results show that the best results can be achieved when A, B and C are 0.4, 0.4 and 0.2. Therefore, this set of parameters is taken as the actual weight values of our network.

Table 2. The results of the loss function using different weighted parameter values. The best performance is shown in bold.

Loss function weights A, B, C	Dice (%)↑	95HD (mm)↓	MPA (%)↑	mIoU (%)↑	TPR (%)↑	ASD (mm)↓
0.2, 0.2, 0.6	91.61	1.37	94.59	85.45	91.05	0.50
0.3, 0.3, 0.4	91.99	1.33	95.41	86.01	92.75	0.48
0.4, 0.4, 0.2	**92.06**	**1.06**	95.91	**86.05**	93.85	**0.47**
0.45, 0.45, 0.1	91.77	1.18	**96.39**	85.61	**94.99**	0.49

3.3.2. Overall performance comparison

In order to verify the segmentation effect of the proposed model, we have quantitatively compared the performance of our proposed CoT-UNet++ with U-Net, UNet++, and TransUNet in terms of Dice, 95HD, MPA, mIoU, TPR, and ASD metrics on our constructed dataset. Three of these comparisons were obtained by running the source code provided by the authors of the published literature, while simply replacing their datasets with our dental dataset. The results of the comparison of the four methods are shown in Table 3, and the best performance is highlighted in bold. From Table 3, our method has the best performance on Dice and 95HD with 92.06% and 1.06 respectively, which can be attributed to the appropriate combination of the CoT block, dense connection, and loss weighting. Compared with the baseline model TransUNet, the Dice evaluation index was improved

by about 5.5%. There was a significant decrease in 95HD, and other indexes were also improved.

Table 3. Quantitative comparison of segmentation performance using different methods on the CBCT dental dataset. (Dice, 95HD, MPA, mIoU, TPR and ASD)

Framework	Dice (%)↑	95HD (mm)↓	MPA (%)↑	mIoU (%)↑	TPR (%)↑	ASD (mm)↓
UNet [13]	86.59	3.46	88.52	83.59	87.87	0.59
UNet++ [20]	86.48	2.14	93.73	83.44	89.84	0.49
TransUNet [31]	89.74	2.60	94.25	82.64	90.99	0.53
Ours(CoT-UNet++)	**92.06**	**1.06**	**95.91**	**86.05**	**93.85**	**0.48**

3.3.3. Ablation experiment

To verify the effectiveness of each component of the proposed method, we conducted the ablation study in this section. First, we qualitatively compared the effectiveness of the CoT block and dense connections under the same loss function. TransUNet, TransUNet++, CoT-UNet, and CoT-UNet++ are selected separately for comparison, as shown in Figure 2. TransUNet++ only uses dense connection on the basis of TransUNet without taking CoT blocks, and the hybrid encoder uses a combination of ResNet-50 and ViT. CoT-UNet does not use dense connections while the CoT block is used, and the hybrid encoder is a combination of CoTNet-50 and ViT. Our proposed CoT-UNet++ utilizes a dense connection, hybrid encoder using a combination of CoTNet-50 and ViT with a weighted loss function. According to Figure 2, we can observe that teeth segmented by CoT-UNet++ can obtain more accurate segmentation results, while several other methods tend to produce under-segment or over-segment results. For the segmentation results of TransUNet and TransUNet++, there is a certain degree of over-segmentation in the molar part. It identifies parts of the alveolar bone as teeth, as well as two adjacent molars that are not well separated from each other. In contrast, the CoT-UNet segmentation results showed incomplete molar segmentation and cavities in the middle of the teeth. Our proposed CoT-UNet++ overcomes these limitations and has better performance.

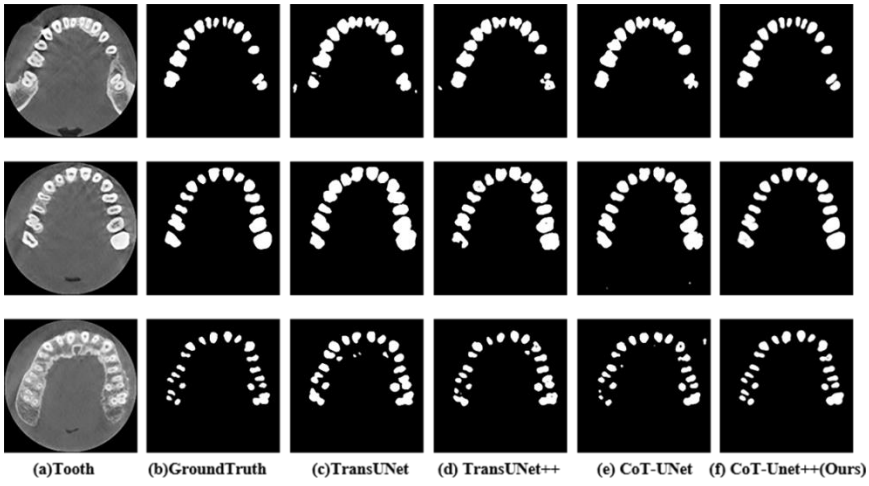

Figure 2. Segmentation results of different CBCT images of teeth with labels. From left to right are the segmentation results using TransUNet, TransUNet++, CoT-UNet and our proposed method respectively.

Performance comparison of 50-layer networks. To qualitatively compare the above segmentation performance and the effectiveness of the weighted loss function, we further compared the performance of our proposed network using different encoders, with or without dense connectivity, with (w) or without (w/o) weighted loss function $\mathcal{L}$. The results are all presented in Table 3. The comparison between TransUNet and TransUNet++ shows that the latter has better values of all evaluation metrics than the former. It shows that the dense connection effectively captures different local features by multi-scale fusion of different features and obtains more accurate location information. CoT-UNet also shows more improvement in segmentation metrics than TransUNet. It is confirmed that unifying context mining and self-attention learning between keys into a single model is an effective way to enhance representational learning and thus facilitate visual recognition. The proposed CoT-UNet++ model obtains the best performance, demonstrating the effectiveness of combining dense connectivity with an efficient encoder. In addition, comparing the data with and without the weighted loss function in the table, it can be seen that the metrics with the weighted loss function are all better than those without, which also shows the necessity of using the weighted loss function.

Table 4. Ablation experiments with our network CoTUNet++ (50 layers) on the CBCT dental dataset. The best performance is shown in bold.

Framework	Dice (%)↑	95HD (mm)↓	MPA (%)↑	mIoU (%)↑	TPR (%)↑	ASD (mm)↓
TransUNet (R50 w/o $\mathcal{L}$)	89.74	2.60	94.25	82.64	90.99	0.53
Ours (TransUNet: R50 w $\mathcal{L}$)	90.78	2.39	94.83	84.17	91.90	0.54
Ours (TransUNet++: R50 + dense w/o $\mathcal{L}$)	91.12	1.55	**96.43**	84.65	**95.33**	0.52
Ours (TransUNet++: R50 + dense w $\mathcal{L}$)	91.43	1.38	95.89	85.11	93.97	0.48
Ours (CoT-UNet: C50 w/o $\mathcal{L}$)	91.30	2.61	94.84	84.93	91.75	0.70
Ours (CoT-UNet: C50 w $\mathcal{L}$)	91.72	1.34	95.54	85.59	93.15	0.50
Ours (CoT-UNet++: C50 + dense w/o $\mathcal{L}$)	91.78	1.18	96.39	85.61	94.99	0.49
Ours (CoT-UNet++: C50 + dense w $\mathcal{L}$)	**92.06**	**1.06**	95.91	**86.05**	93.85	**0.47**

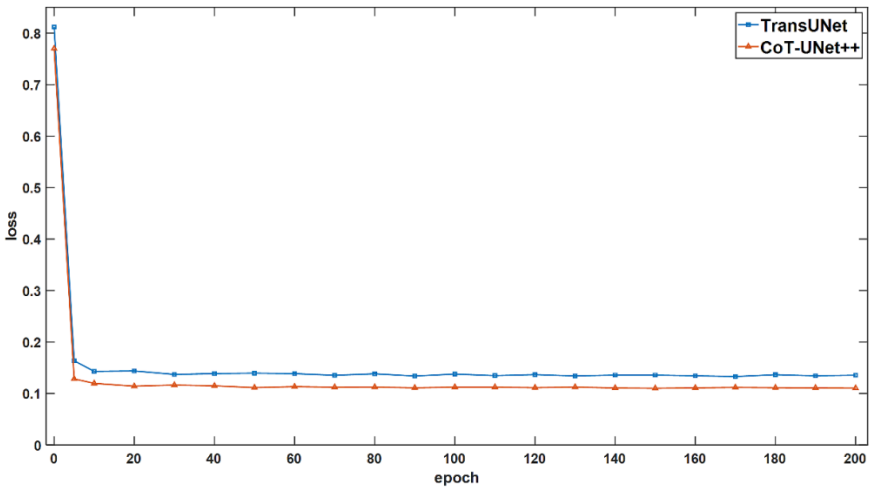

Figure 3. Loss curves of TransUNet and CoT-UNet++.

Figure 3 shows the loss curves of TransUNet and our proposed CoT-UNet++. We can see that the loss value of CoT-UNet++ decreases faster than that of TransUNet, and the final stabilization value is smaller, which can indicate that CoT-UNet++ training is more effective.

Performance comparison of 101-layer networks. To verify the general applicability of our network, we also used ResNet-101 as the baseline model for the encoder. The 3×3 convolution in ResNet-101 is replaced by the CoT block to obtain CoTNet-101. Table 5 compares the performance of several different networks based on ResNet-101 and CoTNet-101. The table shows that the metrics of dental segmentation using 101-layer networks are generally better than those using 50 layers. Moreover, the use of CoT block, dense connections, and weighted loss functions can all improve the accuracy of segmentation, further validating the effectiveness of our proposed network.

Table 5. Ablation experiments with our network CoTUNet++ (101 layers) on the CBCT dental dataset. The best performance is shown in bold.

Framework	Dice (%)↑	95HD (mm)↓	MPA (%)↑	mIoU (%)↑	TPR (%)↑	ASD (mm)↓
TransUNet (R101 w/o $\mathcal{L}$)	89.89	6.65	94.67	82.89	91.90	0.52
Ours (TransUNet: R101 w $\mathcal{L}$)	90.61	2.30	95.20	83.95	92.80	0.58
Ours (TransUNet++: R101 + dense w/o $\mathcal{L}$)	90.98	1.55	96.34	84.44	95.20	0.51
Ours (TransUNet++: R101 + dense w $\mathcal{L}$)	91.86	1.25	95.78	85.78	**95.82**	0.47
Ours (CoT-UNet: C101 w/o $\mathcal{L}$)	91.75	1.35	95.14	85.64	92.22	0.49
Ours(CoT-UNet: C101 w $\mathcal{L}$)	92.11	1.33	95.18	86.20	92.21	0.50
Ours (CoT-UNet++: C101 + dense w/o $\mathcal{L}$)	91.86	2.22	95.03	85.80	91.94	0.59
Ours (CoT-UNet: C101 + dense w $\mathcal{L}$)	**92.27**	**1.17**	**96.35**	**86.41**	94.67	**0.43**

In summary, the segmentation results of our proposed model CoT-UNet++ for dental CBCT images are better than the existing state-of-the-art models and are much closer to the real manual segmentation results from the dentist and physician. However, our model has a large number of parameters and a long training time. Moreover, the segmentation result is not satisfactory for poor-quality CBCT images. Therefore, the network needs to be improved to reduce the number of parameters and to focus more on boundary information.

4. Conclusions

In this paper, we propose a CoT-UNet++ algorithm for dental image segmentation. The CoT block is introduced to enhance visual characterization in the encoder. In order to obtain more accurate localization of teeth in CBCT images, CoTNet-50 and ViT are used as hybrid encoders and the dense connection is utilized to fuse all the same scale feature mappings of the encoder. Moreover, an effective weighted combination of loss functions can capture the tooth structure at the pixel level, making the boundaries of segmented teeth more accurate. Experimental results show that our proposed method achieves better performance in terms of teeth segmentation accuracy over other related methods. However, there is still inaccurate segmentation of individual tooth boundaries, and more attention should be paid to the boundary information in future studies. In addition, tooth classification and 3D tooth reconstruction are worthy of further investigation.

Acknowledgments

This research was funded in part by the National Natural Science Foundation of China under Grant No. 62001267, the Natural Science Foundation of Shandong Province under Grant No. ZR2020QF013, the Fundamental Research Funds for the Central Universities (2022JC017), and the Future Plan for Young Scholars of Shandong University.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. W. R. Proffit, H. W. Fields, D. M. Sarver, *Contemporary orthodontics: Elsevier Health Sciences*, Philadelphia, USA, 2006.
2. P. Holm-Pedersen, M. Vigild, I. Nitschke, D. B. Berkey, Dental care for aging populations in Denmark, Sweden, Norway, United kingdom, and Germany, *J. Dent. Educ.*, **69** (2005), 987–997. http://dx.doi.org/10.1002/j.0022-0337.2005.69.9.tb03995.x
3. C. Tian, Y. Zhang, W. Zuo, C. Lin, D. Zhang, Y. Yuan, A heterogeneous group CNN for image super-resolution, *IEEE Trans. Neural Networks Learn. Syst.*, 2022. https://doi.org/10.1109/TNNLS.2022.3210433
4. C. Tian, Y. Yuan, S. Zhang, C. Lin, W. Zuo, D. Zhang, Image super-resolution with an enhanced group convolutional neural network, *Neural Networks*, **153** (2022). https://doi.org/10.1109/TCSVT.2022.3175959
5. C. Tian, Y. Xu, Z. Li, W. Zuo, L. Fei, H. Liu, Attention-guided CNN for image denoising, *Neural Netw.*, **124** (2020), 117–129. https://doi.org/10.1016/j.neunet.2019.12.024
6. T. Huang, X. Ben, C. Gong, B. Zhang, R. Yan, Q. Wu, Enhanced spatial-temporal salience for cross-view gait recognition, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2022), 6967–6980. https://doi.org/10.1109/TCSVT.2022.3175959
7. X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, W. Meng, Coupled patch alignment for matching cross-view gaits, *IEEE Trans. Image Process.*, **28** (2019), 3142–3157. https://doi.org/10.1109/tip.2019.2894362
8. Y. Guo, B. Li, X. Ben, Y. Ren, J. Zhang, R. Yan, et al., A magnitude and angle combined optical flow feature for microexpression spotting, *IEEE Multimedia*, **28** (2021), 29–39. https://doi.org/10.1109/MMUL.2021.3058017
9. B. Zhang, R. Wang, X. Wang, J. Han, R. Ji, Modulated convolutional Networks, *IEEE Trans/ Neural Netw. Learn. Syst.*, 2021. https://doi.org/10.1109/TNNLS.2021.3060830
10. T. Zhou, J. Si , L. Wang, C. Xu, Automatic detection of underwater small targets using forward-looking sonar images, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022), 1–12. https://doi.org/10.1109/TGRS.2022.3181417
11. C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, N. Sang, BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation, *Int. J. Comput. Vision*, **129** (2021), 3051–3068. https://doi.org/10.1007/s11263-021-01515-2

12. X. Zhong, S. Tu, X. Ma, K. Jiang, W. Huang, Z. Wang, Rainy WCity: A real rainfall dataset with diverse conditions for semantic driving scene understanding, in *International Joint Conference on Artificial Intelligence*, (2022), 1743–1749. https://doi.org/10.24963/ijcai.2022/243
13. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2015), 234–241. https://doi.org /10.1007/978-3-319-24574-4
14. F. Milletari, N. Navab, S. A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in *2016 Fourth International Conference on 3D Vision (3DV)*. (2016), 565–571. https://doi.org/10.1109/3DV.2016.79
15. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2016), 424–432. https://doi.org/10.1007/978-3-319-46723-8_49
16. L. Yu, J. Z. Cheng, Q. Dou, X. Yang, H. Chen, J. Qin, et al., Automatic 3D cardiovascular MR segmentation with densely-connected volumetric convnets, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2017), 287–295. https://doi.org/10.48550/arXiv.1708.00573
17. X. Li, H. Chen, X. Qi, Q. Dou, C. W. Fu, P. Heng, H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes, *IEEE Trans. Med. Imaging*, **37** (2018), 2663–2674. https://doi.org/10.1109/tmi.2018.2845918
18. Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, A. L. Yuille, Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 8280–8289. https://doi.org/10.1109/CVPR.2018.00864
19. Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, A. L. Yuille, A fixed-point model for pancreas segmentation in abdominal CT scans, in *International Conference on Medical Image Computing and Computer-assisted Intervention*, (2017), 693–701. http://dx.doi.org/10.1007/978-3-319-66182-7_79
20. Z. Zhou, M. M. Rahman 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*, (2018), 3–11. https://doi.org/10.1007/978-3-030-00889-5_1
21. 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2020), 1055–1059. https://doi.org/10.1109/ICASSP40776.2020.9053405
22. J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, V. M. Patel, Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2020), 363–373. https://doi.org/10.1007/978-3-030-59719-1_36
23. J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, et al., Attention gated networks: Learning to leverage salient regions in medical images, *Med. Image Anal.*, **53** (2019), 197–207. https://doi.org/10.1016/j.media.2019.01.012

24. X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 7794–7803. https://doi.org/10.1109/CVPR.2018.00813
25. H. Akhoondali, R. A. Zoroofi, G. Shirani, Rapid automatic segmentation and visualization of teeth in ct-scan data, *J. Appl. Sci.*, **9** (2019), 2031–2044. https://dx.doi.org/10.3923/jas.2009.2031.2044
26. D. X. Ji, S. H. Ong, K. W. C. Foong, A level-set based approach for anterior teeth segmentation in cone beam computed tomography images, *Comput. Biol. Med.*, **50** (2014), 116–128. https://doi.org/10.1016/j.compbiomed.2014.04.006
27. Y. Gan, Z. Xia, J. Xiong, Q. Zhao, Y. Hu, J. Zhang, Toward accurate tooth segmentation from computed tomography images using a hybrid level set model, *Med. Phys.*, **42** (2015), 14–27. https://doi.org/10.1118/1.4901521
28. Y. Pei, X. Ai, H. Zha, T. Xu, G. Ma, 3d exemplar-based random walks for tooth segmentation from cone-beam computed tomography images, *Med. Phys.*, **43** (2016), 5040–5050. https://doi.org/10.1118/1.4960364
29. S. Barone, A. Paoli, A. V. Razionale, Ct segmentation of dental shapes by anatomy-driven reformation imaging and b-spline modelling, *Int. J. Numer. Methods Biomed. Eng.*, **32** (2016), e02747. https://doi.org/10.1002/cnm.2747
30. Z. Cui, C. Li, W. Wang, ToothNet: Automatic tooth instance segmentation and identification from cone beam CT images, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 6361–6370. https://doi.org/10.1109/CVPR.2019.00653
31. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, et al., Transunet: Transformers make strong encoders for medical image segmentation, preprint, arXiv:2102/04306.
32. I. Bello, B. Zoph, A. Vaswani, J. Shlens, Q. V. Le, Attention augmented convolutional networks, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 3286–3295. https://doi.org/10.1109/ICCV.2019.00338
33. 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*, (2020), 213–229. https://doi.org/10.1007/978-3-030-58452-8_13
34. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, An image is worth 16x16 words: Transformers for image recognition at scale, preprint, arXiv:2010/11929.
35. Y. Li, Y. Pan, T. Yao, J. Chen, T. Mei, Scheduled sampling in vision-language pretraining with decoupled encoder-decoder network, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 8518–8526. http://dx.doi.org/10.1609/aaai.v35i10.17034
36. Y. Pan, T. Yao, Y. Li, T. Mei, X-linear attention networks for image captioning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 10971–10980. https://doi.org/10.1109/CVPR42600.2020.01098
37. P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-alone self-attention in vision models, *Adv. Neural Inform. Process. Syst.*, **32** (2019), https://doi.org/10.48550/arXiv.1906.05909
38. H. Zhao, J. Jia, V. Koltun, Exploring self-attention for image recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 10076–10085. https://doi.org/10.1109/CVPR42600.2020.01009

39. Y. Li, T. Yao, Y. Pan, T. Mei, Contextual transformer networks for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. https://doi.org/10.48550/arXiv.2107.12292
40. T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 2980–2988. https://doi.org/10.1109/ICCV.2017.324
41. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

AIMS Press