Mathematical Biosciences and Engineering

MBE, 20(10): 18650–18669.
DOI: 10.3934/mbe.2023827
Received: 12 July 2023
Revised: 13 September 2023
Accepted: 18 September 2023
Published: 07 October 2023

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

Research article

ECA-TFUnet: A U-shaped CNN-Transformer network with efficient channel attention for organ segmentation in anatomical sectional images of canines

Yunling Liu[1,*], Yaxiong Liu[1], Jingsong Li[1], Yaoxing Chen[2], Fengjuan Xu[3], Yifa Xu[4], Jing Cao[2] and Yuntao Ma[5]

[1] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
[2] College of Veterinary Medicine, China Agricultural University, Beijing 100193, China
[3] Animal and Plant Disease Prevention and Control Center, Chaoyang District, Beijing 100016, China
[4] Shandong Digihuman Technology Co., Ltd, Shandong 250100, China
[5] College of Land Science and Technology, China Agricultural University, Beijing 100193, China

* **Correspondence:** Email: liuyunling@cau.edu.cn; Tel: +8613718631215.

Abstract: Automated organ segmentation in anatomical sectional images of canines is crucial for clinical applications and the study of sectional anatomy. The manual delineation of organ boundaries by experts is a time-consuming and laborious task. However, semi-automatic segmentation methods have shown low segmentation accuracy. Deep learning-based CNN models lack the ability to establish long-range dependencies, leading to limited segmentation performance. Although Transformer-based models excel at establishing long-range dependencies, they face a limitation in capturing local detail information. To address these challenges, we propose a novel ECA-TFUnet model for organ segmentation in anatomical sectional images of canines. ECA-TFUnet model is a U-shaped CNN-Transformer network with Efficient Channel Attention, which fully combines the strengths of the Unet network and Transformer block. Specifically, The U-Net network is excellent at capturing detailed local information. The Transformer block is equipped in the first skip connection layer of the Unet network to effectively learn the global dependencies of different regions, which improves the representation ability of the model. Additionally, the Efficient Channel Attention Block is introduced to the Unet network to focus on more important channel information, further improving the robustness of the model. Furthermore, the mixed loss strategy is incorporated to alleviate the problem of class imbalance. Experimental results showed that the ECA-TFUnet model yielded 92.63%

IoU, outperforming 11 state-of-the-art methods. To comprehensively evaluate the model performance, we also conducted experiments on a public dataset, which achieved 87.93% IoU, still superior to 11 state-of-the-art methods. Finally, we explored the use of a transfer learning strategy to provide good initialization parameters for the ECA-TFUnet model. We demonstrated that the ECA-TFUnet model exhibits superior segmentation performance on anatomical sectional images of canines, which has the potential for application in medical clinical diagnosis.

Keywords: anatomical sectional images of canines; segmentation; transformer; efficient channel attention; Unet network; transfer learning

1. Introduction

Organ segmentation from anatomical sectional images is a key component of clinical applications [1], as well as a critical step in the 3D reconstruction of organs [2]. Accurate organ segmentation from anatomical sectional images of canines can help veterinarians precisely identify the class and shape of organs, providing reliable assistance in clinical diagnosis and treatment. Besides, it also provides a wealth of material for the education of animal clinical medicine and canine anatomy.

Traditional segmentation methods rely on manual delineation by experts. For instance, Park et al. [3] manually delineated organ boundaries, such as the heart and lungs, in anatomical sectional images of canines. However, this method is time-consuming and non-reproducible. Existing semi-automatic segmentation methods based on image processing, such as threshold segmentation, edge detection, active contour method and level set, have also been employed to handle organ segmentation tasks. JSeo Park et al. [4] used a threshold segmentation method to segment organs and tissue structures in anatomical sectional images of canines. Czeibert et al. [5] used Amira software for semi-automatic segmentation of the brain, bones, arteries and veins in anatomical sectional images of canines, which is crucial for 3D organ reconstruction. Xiu Shu et al. [6] used an improved active contour model to segment cardiac MR images with intensity inhomogeneity and achieved good results. Furthermore, they employed an adaptive local variances-based level set [7] to segment medical images affected by intensity inhomogeneity and noise, including the cardiac MR, brain MR and breast ultrasound images. Although semi-automatic segmentation methods based on image processing have higher efficiency than manual delineation, they tend to result in lower generalization performance. Furthermore, their reliance on a priori knowledge diminishes their level of automation. Therefore, efficient and accurate automated image segmentation has become an urgent demand for the current analysis of anatomical sectional images.

With the rapid development of computer vision technology, deep learning has been widely used in various medical image analysis tasks with remarkable success [8]. CNNs are one of the most commonly used models and have the ability to automate image segmentation [9]. Several CNN models, such as Full Connected Network (FCN) [10], DenseNet [11], Deeplabv3+ [12] and Unet [13], have been successfully employed in the domain of medical image segmentation. Notably, Unet is the first CNN model to be applied to medical image segmentation and demonstrate exceptional performance. The skip connection structure of Unet fuses deep and shallow features to reduce information loss, resulting in more precise segmentation outcomes. Schmid et al. [14] used the Unet model to segment the medial retropharyngeal lymph nodes of the canine in tomographic images. Park et al. [15] used the

fully convolutional DenseNet model to segment the organs of the canine head and neck in tomographic images. However, CNN models suffer from the inherent limitations in convolutional operations [16], which result in difficulty in accurately capturing global contextual information and establishing long-range dependencies. The ability to construct global contextual information is essential for intensive prediction tasks during medical image segmentation, either within a single medical image or between adjacent medical images [17].

Recently, the success of the Transformer, which can capture long-range dependencies, has the potential ability to solve the above problems. The Transformer is a successful example of applying the way of processing sequence data in natural language processing to the field of computer vision and performs well in tasks such as image recognition [18], image detection [19] and image segmentation [20]. Dosovitskiy et al. [18] applied a Transformer to the field of computer vision and proposed the Vision Transformer (ViT) model, which was used for the medical image classification task. TransUNet [21] is the first medical image segmentation network based on the transformer with excellent segmentation results. In contrast to the CNN models, the Transformer relies on the self-attention mechanism to model long-range sequential dependencies and it excels in global feature modeling and exhibits great transferability [22]. Furthermore, the transformer can mitigate the impact of shallow features on overall network performance through skip connection [23]. Although Transformer models excel at capturing global contextual information, they lack the ability to get local detail information [21].

In this work, we propose the ECA-TFUnet model, which combines CNN and Transformer block [18], leveraging the strengths of both. The model incorporates the Transformer block into the first skip connection of the Unet network and introduces the Efficient Channel Attention (ECA) block [24] in Unet. Moreover, the mixed loss strategy is adopted to alleviate the class imbalance problem. The ECA-TFUnet model is employed to achieve precise segmentation of 11 organs in anatomical sectional images of canines, offering reliable assistance for clinical diagnosis and anatomical research in canines. Furthermore, to comprehensively evaluate the performance of the model, we also conducted experiments on a public dataset called Combined Healthy Abdominal Organ Segmentation (CHAOS) [25].

The contributions of this article can be summarized as follows:

1) We proposed the ECA-TFUnet model for precise organ segmentation in anatomical sectional images of canines, offering a novel idea for the combination of CNN and Transformer.

2) To comprehensively evaluate the performance of the ECA-TFUnet model, we compared it with 11 state-of-the-art models and conducted experiments on the CHAOS dataset.

3) We designed a transfer learning strategy using the CHAOS dataset as the source data to further improve the performance of the ECA-TFUnet model.

2. Materials and methods

2.1. Acquisition and preprocessing of sectional anatomical images of canines

The dataset was generously provided by Laboratory of Anatomy of Domestic Animal, National Key Laboratory of Veterinary Public Health and Safety, College of Veterinary Medicine, China Agricultural University. Teledyne DALSA Piranha XL 16K camera and Schneider-KREUZNACH Apo-Componon 4.5/90 lens were used to take anatomical sectional images of the thoracoabdominal region of the beagle. The image resolution was 16,384 × 38,000 pixels, and a total of 500 anatomical

sectional images were recorded, as shown in Figure 1(a). To speed up the convergence rate of the model, the ice background area was cropped and replaced with a black background, shown in Figure 1(b). To save training and inference time, the resolution of the images was uniformly adjusted to 256 × 256 pixels and converted to grayscale images. The preprocessed images with 11 organs are shown in Figure 2. Labelme program (https://github.com/wkentaro/labelme) was used to label these 11 organs in images. The dataset was divided into the training set, validation set and test set according to the ratio of 7:2:1. During experimental training, we employed five frequently used data augmentation techniques to enhance the diversity of the training dataset. These techniques included random rotation (-10°–10°), random resizing (scale factor 0.9), vertical deformation (magnitude = 0.1), perspective deformation (magnitude = 0.1) and elastic deformation (magnitude = 4).

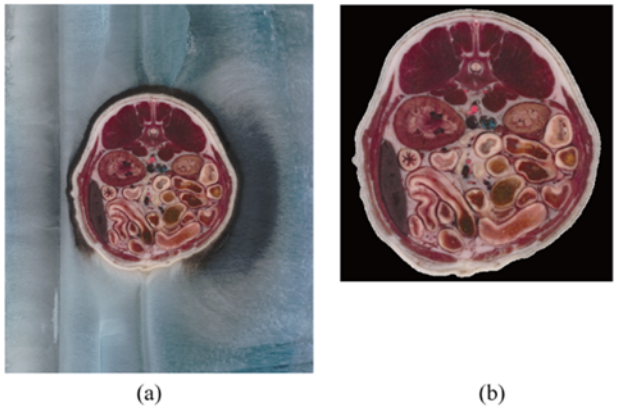

Figure 1. Visualization of the initial preprocessing results. (a) original image (b) after initial preprocessing.

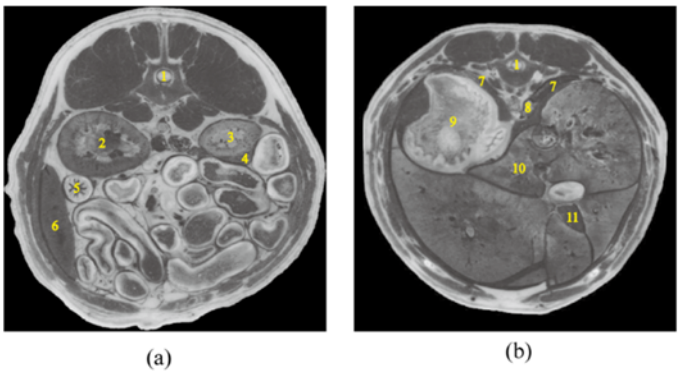

Figure 2. (a) and (b) show the distribution of organs after the last step of image preprocessing labeled as follows: 1. Spinal Cord; 2. Kidney (L); 3. Kidney (R); 4. Pancreas; 5. Intestine; 6. Spleen; 7. Septum; 8. Lung; 9. Stomach; 10. Liver; 11. Gallbladder.

2.2. Preprocessing of the CHAOS dataset

The Combined Healthy Abdominal Organ Segmentation-T1DUAL in phase MRI (CHAOS) dataset [25] was utilized as our experimental data. This dataset consists of 647 MRI images acquired from healthy individuals and includes four abdominal organs: liver, left kidney, right kidney and spleen.

To eliminate interference from irrelevant regions and expedite the convergence of the model, we cropped out most of the irrelevant black background area, adjusted the image brightness and contrast and standardized the image resolution to 256 × 256 pixels, as demonstrated in Figure 3. In the experiments, the dataset was also randomly divided into training, validation and test sets with a ratio of 7:2:1, and the same data augmentation operations as Section 2.1 were performed during training.

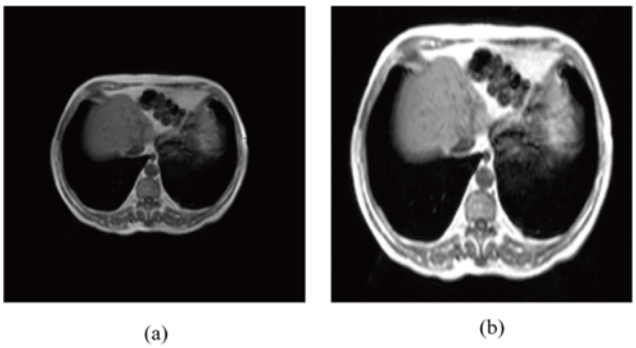

Figure 3. Image preprocessing visualization of CHAOS dataset. (a) original image (b) after pre-processing.

2.3. Overview

Figure 4 shows the flowchart of anatomical sectional images of the segmentation method. The ECA-TFUnet model adopts a U-shaped encoder-decoder framework with skip connections. Specifically, the Transformer block was integrated into the first skip connection, and the ECA block was incorporated into the encoder-decoder framework. Additionally, the mixed loss strategy was employed to further enhance the performance of the model. Finally, the performance of the trained ECA-TFUnet model was evaluated through evaluation metrics.

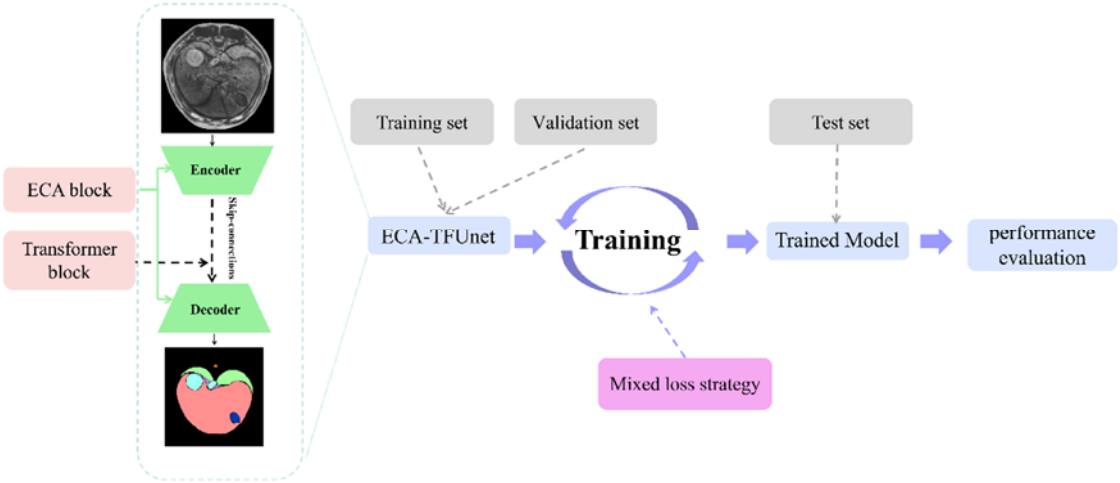

Figure 4. The flowchart of sectional anatomical images of the segmentation method.

2.4. ECA-TFUnet architecture

The architecture of ECA-TFUnet is illustrated in Figure 5. It consists of three major parts: the encoder, the decoder and skip connections, which is the basic structure of Unet. The encoder contains the CNN network with a ResNet50 backbone [26] and the ECA block. The features of images are extracted by the CNN network, and then the obtained feature maps are input into the ECA block to weigh the channels of the feature maps, focusing on the more important feature channels. In the decoder part, the low-resolution features extracted from the encoder are recovered to the full resolution of the input image by cascading multiple upsamplers. The ECA block is also used in the decoder part to enhance important feature channels and suppress irrelevant ones. Skip connections are utilized to fuse shallow features with deep features, resulting in richer semantic information. Moreover, the first skip connection uses the Transformer block to establish remote correlations between different local regions of the feature map.

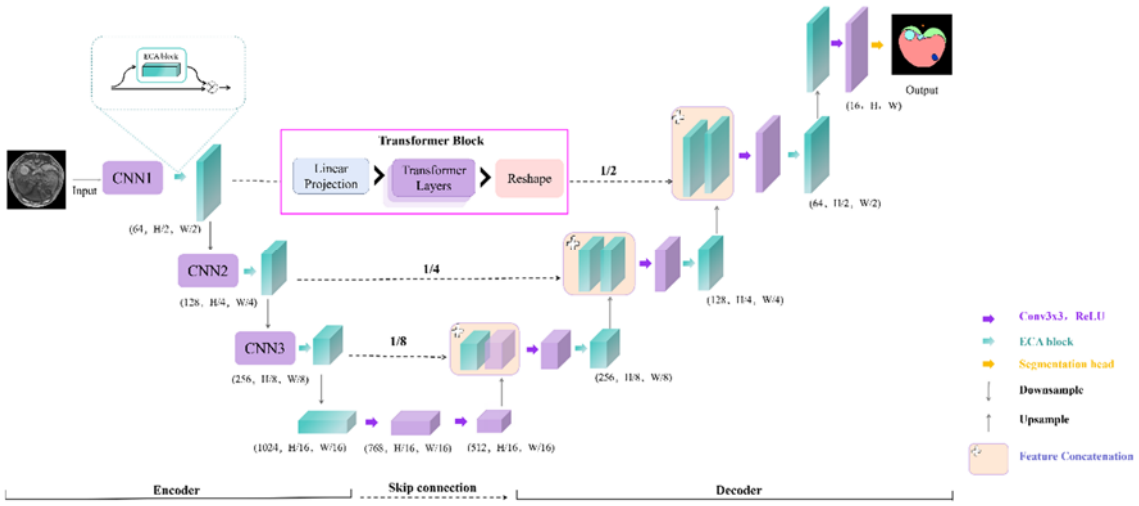

Figure 5. The overall architecture of ECA-TFUnet.

2.5. Transformer block

The structure of the Transformer block is shown in Figure 6. First, the feature map is operated by Image Sequentialization, which can reshape the feature map into a sequence of 2-dimensional patches. Then, these patches are mapped to a latent D-dimensional embedding space through the trainable linear projection layer. In addition, the position embeddings are added to the patch embeddings to ensure that each patch has the correct spatial position relationship. The formula is shown in Eq (1).

$$Z_0 = [X_p^1\mathrm{E}; X_p^2\mathrm{E}; \ldots X_p^N\mathrm{E}] + E_{pos} \tag{1}$$

where Z_0 denotes the final vectorized patches inputted into the transformer layer. X_p^1 to X_p^N denotes the vectorized patches from 1 to N, N denotes the number of patches and p denotes the size of patches. $\mathrm{E} \in R^{(p^2 \times C) \times D}$ denotes patch projection, C denotes the number of channels and $E_{pos} \in R^{N \times D}$ denotes

position embedding.

The Transformer block contains $\ell = 12$ layers, and each transformer layer contains a multi-head attention and a multilayer perceptron. Z_0 is input into the transformer layer for training and the training output of ℓth layer can be acquired by Eq (2).

$$\hat{Z}_\ell = MSA\big(LN(Z_\ell - 1)\big) + Z_\ell - 1$$

$$Z_\ell = MLP\left(LN\left(\hat{Z}_\ell\right)\right) + \hat{Z}_\ell \tag{2}$$

where $MSA(.)$ denotes multi-head attention [27]. $MLP(.)$ denotes the multilayer perceptron, $LN(.)$ denotes the layer normalization and Z_ℓ denotes the encoded feature representation. Multi-head attention can focus on the global contextual information to solve the long-distance dependency problem. First, MSA projects queries, keys and values by using learnable linear layers. Then, these projected groups are fed into the Scaled Dot-Product Attention module for parallel processing. Finally, the resulting outputs are concatenated and passed into a multilayer perceptron, as depicted in Eq (3). Multilayer perceptron can analyze its inter-patch dependencies and aggregate information to finalize the prediction task.

$$\mathrm{MSA(Q,K,V)} = Concat(head_1, head_2, \ldots, head_h)W^O \tag{3}$$

$$\text{where } head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

where Q, K and V denote query, key and value respectively. W_i^Q, W_i^K, W_i^V denote the learnable linear matrices of Q, K and V respectively. Attention(.) denotes the Scaled Dot-Product Attention module which can be acquired by Eq (4):

$$\mathrm{Attention(Q,K,V)} = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{4}$$

where d_k denotes the dimension of Q and K.

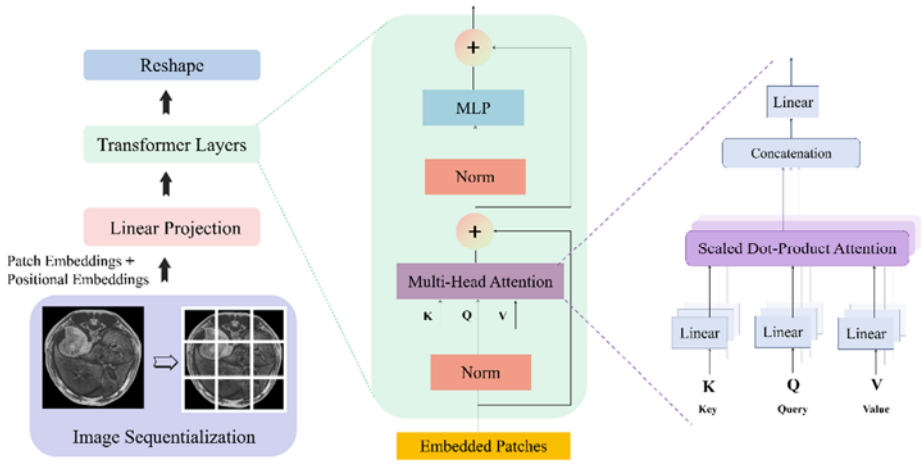

Figure 6. The overall architecture of the Transformer block.

2.6. ECA block

The structure of the ECA block is shown in Figure 7. Firstly, to aggregate the channel information of the feature map, global average pooling is performed on the feature map $Y \in R^{(W\times H\times C)}$ to obtain the vector $Y_{avg} \in R^{(1\times 1\times C)}$ which is expressed by Eq (5). Then, the one-dimensional convolution of the vector Y_{avg} is performed to complete cross-channel threshold interaction to obtain the post-interaction weights W which can be acquired by Eq (6). Finally, the weights are weighted into the original tensor to obtain the new tensor.

$$Y_{avg} = \mathrm{GAP(Y)} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} Y_{i,j} \tag{5}$$

where GAP(.) denotes global average pooling. Y denotes the input feature map. H and W denote the length and width of the feature map, respectively.

$$\mathrm{W} = \sigma(\mathrm{C1}D_k(\mathrm{Y})) \tag{6}$$

where σ denotes the sigmoid activation function. C1D denotes the one-dimensional convolution. k denotes the size of the convolution kernel and an adaptive adjustment strategy is used to assign a value to k, which is given in Eq (7).

$$\mathrm{C} = \phi(\mathrm{k}) = 2^{(y*k-b)} \tag{7}$$

where C denotes the feature map channel size. The experiments in this paper set the γ and b parameters to 2 and 1, respectively, and take the logarithm of the left and right sides of Eq (7) simultaneously to obtain the convolution kernel k, whose formula is expressed by Eq (8).

$$\mathrm{k} = \varphi(\mathrm{C}) = \left|\frac{\log_2(C)}{y} + \frac{b}{y}\right|_{odd} \tag{8}$$

where $|t|_{odd}$ denotes the nearest odd number of t.

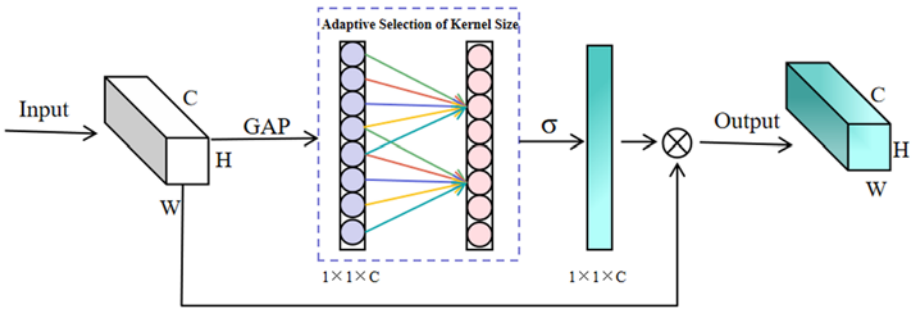

Figure 7. The overall architecture of the ECA block.

2.7. Mixed loss strategy

The mixed loss strategy used in ECA-TFUnet is defined as Eq (9).

$$\text{Loss} = \alpha L_{CE} + (1-\alpha) L_{DICE} \tag{9}$$

where L_{CE} denotes cross-entropy loss, which is used to evaluate the accuracy of the average predicted pixel, and is defined by Eq (10). L_{DICE} denotes dice loss, which can be expressed by Eq (11). α denotes the weighting factor with a value range between 0 and 1, and is used to adjust the weight of the L_{CE} and L_{DICE}.

$$L_{CE} = -\sum_i^N g_i * \log \frac{e^{p_i}}{\sum_i^N e^{p_i}} \tag{10}$$

where *N* denotes the total number of pixels. g_i denotes the i-th pixel point in the ground truth image. p_i denotes the i-th pixel point of the predicted result. However, L_{CE} is weak in dealing with the category imbalance problem. When the number of pixels between categories differs significantly, it degrades the performance of the model for the segmentation of categories with fewer pixels. To solve this problem, a second loss, which is L_{DICE}, is added to this model.

$$L_{DICE} = 1 - \frac{2*\sum_i^N p_i * g_i}{\sum_i^N p_i + \sum_i^N g_i} \tag{11}$$

where *N*, g_i and p_i have the same meanings as indicated in Eq (10).

2.8. Evaluation metrics

Intersection over Union (IoU), Dice Similarity Coefficient (DSC) and Accuracy (ACC) metrics are applied to evaluate the performance of the model. The corresponding equations are shown in Eqs (12)–(14).

$$\text{IoU} = \frac{TP}{FN+TP+FP} \tag{12}$$

$$\text{DSC} = \frac{2TP}{2TP+FP+FN} \tag{13}$$

$$\text{ACC} = \frac{TP+TN}{TP+FP+FN+TN} \tag{14}$$

where TP (true positive) denotes the number of samples where both the actual label and the predicted label are positive. FP (false positive) denotes the number of samples where the predicted label is positive and the true label is negative. FN (false negative) denotes the number of samples where the predicted label is negative and the actual label is positive and TN (true negative) denotes the number of samples where both the predicted label and the actual label are negative.

3. Results

3.1. Experimental settings

All the algorithms were performed on a workspace with NVIDIA GeForce GTX 3090Ti GPU equipped with Ubuntu 18.04 LTS 64-bit system. Python 3.7 and the deep learning framework pytorch1.8.0 were used. The proposed model was trained by Adam optimizer with a momentum of 0.9. The batch size

was set to 8, the initial learning rate was 0.01 and the training process consisted of 120 epochs.

3.2. Mixed loss strategy weight ratio analysis

To obtain the optimal weight ratio α of the mixed loss, it was set to 0 (dice loss only), 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1 (cross-entropy loss only), respectively, for testing. The experimental results were shown in Figure 8. It can be seen that α = 0.1 is the best option, where IoU (avg) and DSC (avg) are the highest values. The values of IoU (avg), DSC (avg) and ACC (avg) were 92.63%, 96.07% and 96.39%, respectively.

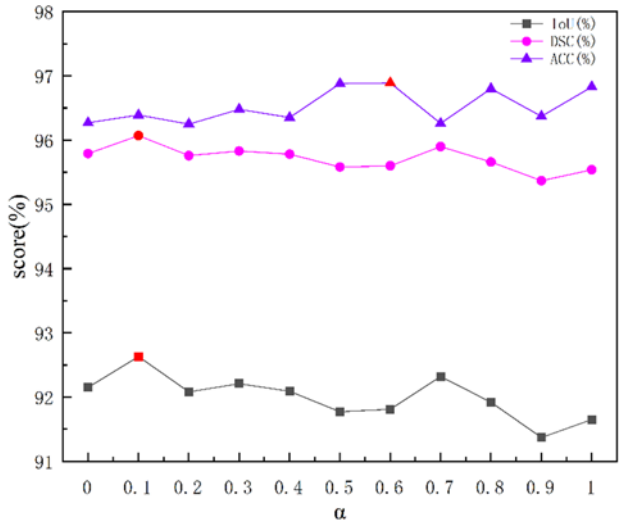

Figure 8. Mixing loss strategy weight ratio experimental results. The red symbol denotes the highest value of the evaluation index.

3.3. Ablation experiments

Table 1. Ablation experiments results.

Transformer block	ECA block	Mixing loss strategy	IoU (avg)%	DSC (avg)%	ACC (avg)%
—	—	—	89.63	94.34	93.41
√	—	—	91.55	95.46	95.43
√	√	—	92.15	95.79	96.27
√	√	√	92.63	96.07	96.39

Ablation experiments were applied to verify the effectiveness of the Transformer block, ECA block and mixed loss strategy in ECA-TFUnet. We added these three parts in sequence without changing the rest of the ECA-TFUnet structure to explore how the added parts will affect the performance of the model. The results are shown in Table 1. It is clear that all these three parts were effective, and any one of them led to an increase in the scores of IoU (avg), DSC (avg) and ACC (avg). The addition of these three parts increased the values of IoU (avg), DSC (avg) and ACC (avg) by 3%, 1.73% and 2.98% in total, respectively.

3.4. Experimental result based on anatomical sectional images of canines

The ECA-TFUnet model was used to segment 11 organs in anatomical sectional images of canines, and the segmentation results were shown in Table 2. The values of IoU (avg), DSC (avg) and ACC (avg) were 92.63%, 96.07% and 96.39%, respectively. To verify the superiority of ECA-TFUnet, we compared it with 11 state-of-the-art models, and the segmentation results are shown in Table 2. It is obvious that the ECA-TFUnet model achieved the best results in all metrics. The top 5 models with higher IoU are ECA-TFUnet (IoU = 92.63%), TransUnet (IoU = 90.96%, 1.67% lower than ECA-TFUnet), Segformer (IoU = 90.50%, 2.13% lower than ECA-TFUnet), Swin-Transformer (IoU = 90.38%, 2.25% lower than ECA-TFUnet) and DeepLabv3+ (IoU = 89.82%, 2.81% lower than ECA-TFUnet). Table 3 shows the organ segmentation results of these top 5 models. The results indicate that the ECA-TFUnet model outperformed all other models in terms of segmentation accuracy for 10 organs.

Table 2. Segmentation results of 12 methods. The boldfaced words in the method column denote the top 5 methods with high IoU (avg). The boldfaced words in the IoU (avg), DSC (avg) and ACC (avg) columns denote the highest value of the corresponding evaluation index.

Method	IoU (avg)%	DSC (avg)%	ACC (avg)%
FCN [10]	88.06	93.29	92.93
PSPNet [28]	88.12	93.32	92.82
DeepLabV3+ [12]	89.82	94.43	94.02
UperNet [12]	89.74	94.39	93.77
DANet [29]	88.59	93.62	93.46
GCNet [30]	88.03	93.27	93.22
OCRNet [31]	89.20	94.05	93.38
ViT [18]	88.99	93.92	93.12
Swin-Transformer [32]	90.38	94.76	94.85
Segformer [33]	90.50	94.83	94.65
TransUnet [21]	90.96	95.15	95.64
ECA-TFUnet	**92.63**	**96.07**	**96.39**

Table 3. The IoU results of the 11 organs for the top 5 models. The boldfaced words denote the highest value of the corresponding evaluation index.

Method	Stomach	Liver	Gallbladder	Lung	Spinal Cord	Septum	Kidney (L)	Kidney (R)	Pancreas	Spleen	Intestine
DeepLabV3+	95.77	97.77	88.63	92.49	78.73	69.98	96.01	94.55	83.88	91.96	89.67
Swin-Transformer	95.76	97.74	88.88	92.26	79.08	70.97	96.33	94.68	86.32	93.10	90.97
Segformer	96.08	97.82	89.15	92.42	79.32	71.41	96.19	94.67	86.58	93.09	90.82
TransUnet	94.05	96.44	89.43	89.46	**87.80**	73.97	95.80	94.55	86.12	92.95	92.52
ECA-TFUnet	**96.59**	**98.06**	**90.41**	**93.74**	86.11	**77.34**	**97.16**	**95.96**	**89.41**	**94.43**	**93.70**

Figure 9 shows the segmentation visualization results of the top 5 models. From the figure, it can be seen that the Segformer, Swin-Transformer and Deeplabv3+ models could not accurately identify the boundary of the septum in row (4). The TransUnet, Segformer, Swin-Transformer and Deeplabv3+ models perform poorly for the gallbladder segmentation in row (5). The ECA-TFUnet model performs better in segmenting 11 organs and has a higher similarity to the label.

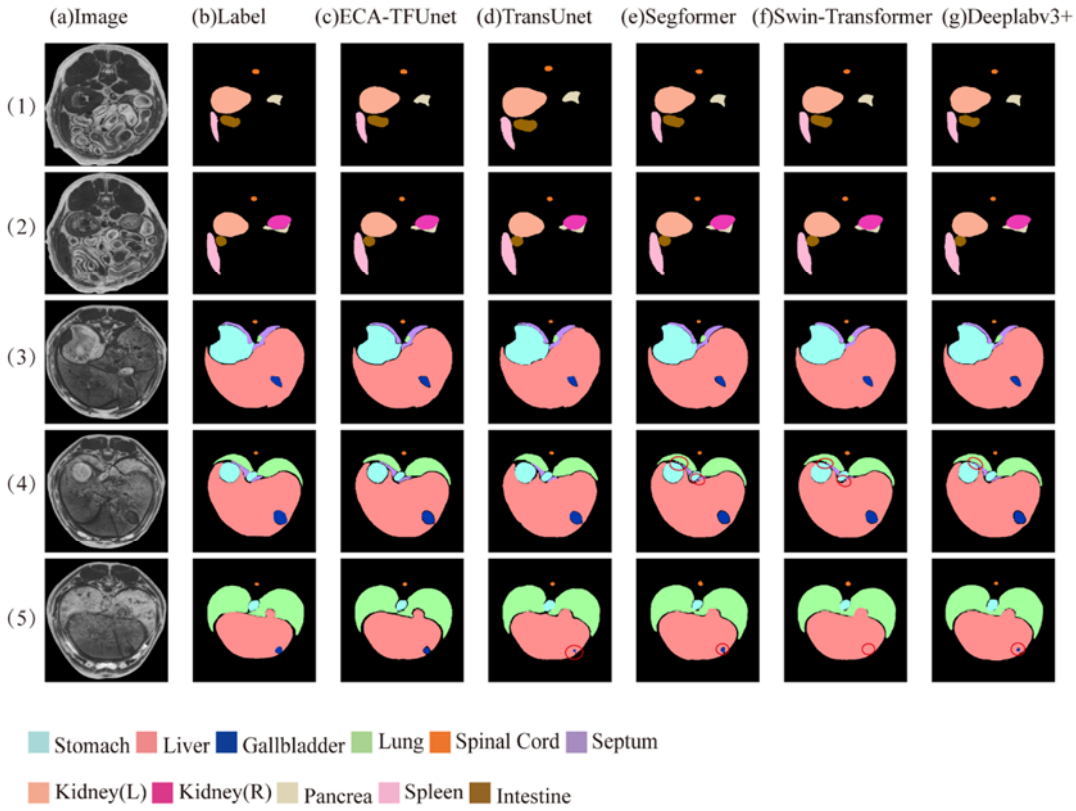

Figure 9. The segmentation visualization results of the top 5 models. The images in rows (1)–(5) were randomly sampled from the anatomical sectional images of canines (including 11 different organs). column (a) is the original image, column (b) is the label image, column (c) is the segmentation result of the ECA-TFUnet and columns (d)–(g) are the results of segmentation by TransUnet, Segformer, Swin-Transformer and Deeplabv3+ models. The red circle is the place of the segmentation error.

3.5. Experiments on CHAOS Dataset

Compared to anatomical sectional images, the CHAOS dataset has blurrier organ boundaries, smaller grayscale differences and more challenging segmentation tasks. To comprehensively evaluate the performance of the ECA-TFUnet model, we also conducted organ segmentation experiments on the CHAOS dataset and still compared it with the 11 state-of-the-art methods. The results in Table 4 show that the ECA-TFUnet get the highest scores of IoU (avg), DSC (avg) and ACC (avg) which were 87.93%, 93.46% and 94.78%, respectively. The top 5 models with higher IoU are ECA-TFUnet (IoU = 87.93%), TransUnet (IoU = 85.83%, 2.1% lower than ECA-TFUnet), Segformer (IoU = 85.27%, 2.66% lower than ECA-TFUnet), Swin-Transformer (IoU = 85.12%, 2.81% lower

than ECA-TFUnet) and DeepLabv3+ (IoU = 85.02%, 2.91% lower than ECA-TFUnet). Table 5 presents the organ segmentation results of the top 5 models on the CHAOS dataset, revealing that ECA-TFUnet achieved the highest scores for all organs, further demonstrating the superiority of ECA-TFUnet.

Table 4. Segmentation results of 12 methods on the CHAOS dataset. The boldfaced words in the method column denote the top 5 methods with high IoU (avg). The boldfaced words in the IoU (avg), DSC (avg) and ACC (avg) columns denote the highest value of the corresponding evaluation index.

Method	IoU (avg)%	DSC (avg)%	ACC (avg)%
FCN	84.37	91.31	90.16
PSPNet	84.33	91.28	89.71
DeepLabV3+	85.02	91.71	90.66
UperNet	84.17	91.18	89.79
DANet	84.69	91.50	90.32
GCNet	84.63	91.47	90.21
OCRNet	83.39	90.69	88.97
ViT	84.86	91.61	90.26
Swin-Transformer	85.12	91.77	90.45
Segformer	85.27	91.86	90.55
TransUnet	85.83	92.21	94.24
ECA-TFUnet	**87.93**	**93.46**	**94.78**

Table 5. The IoU results of the 4 organs for the top 5 models. The boldfaced words denote the highest value of the corresponding evaluation index.

Method	Liver	Kidney (R)	Kidney (L)	Spleen
DeepLabV3+	89.03	78.10	79.72	79.31
Swin-Transformer	88.88	77.66	80.29	79.86
Segformer	89.09	78.39	79.68	80.25
TransUnet	88.42	81.17	82.28	78.35
ECA-TFUnet	**90.29**	**83.86**	**83.88**	**82.50**

Figure 10 shows the segmentation visualization results of the top 5 models. It can be seen that the ECA-TFUnet model could accurately segment the liver edge of the original image in row (2), which was better than the other 4 models. All models had Under-segmentation results of the liver in row (4). Additionally, the Swin-Transformer and Deeplabv3+ models also exhibited inadequate segmentation of the spleen. The Segformer, Swin-Transformer and Deeplabv3+ models could not sufficiently mine the features of the liver in row (5), resulting in the loss of subtle features. In general, the segmentation effect of ECA-TFUnet was significantly better than other models.

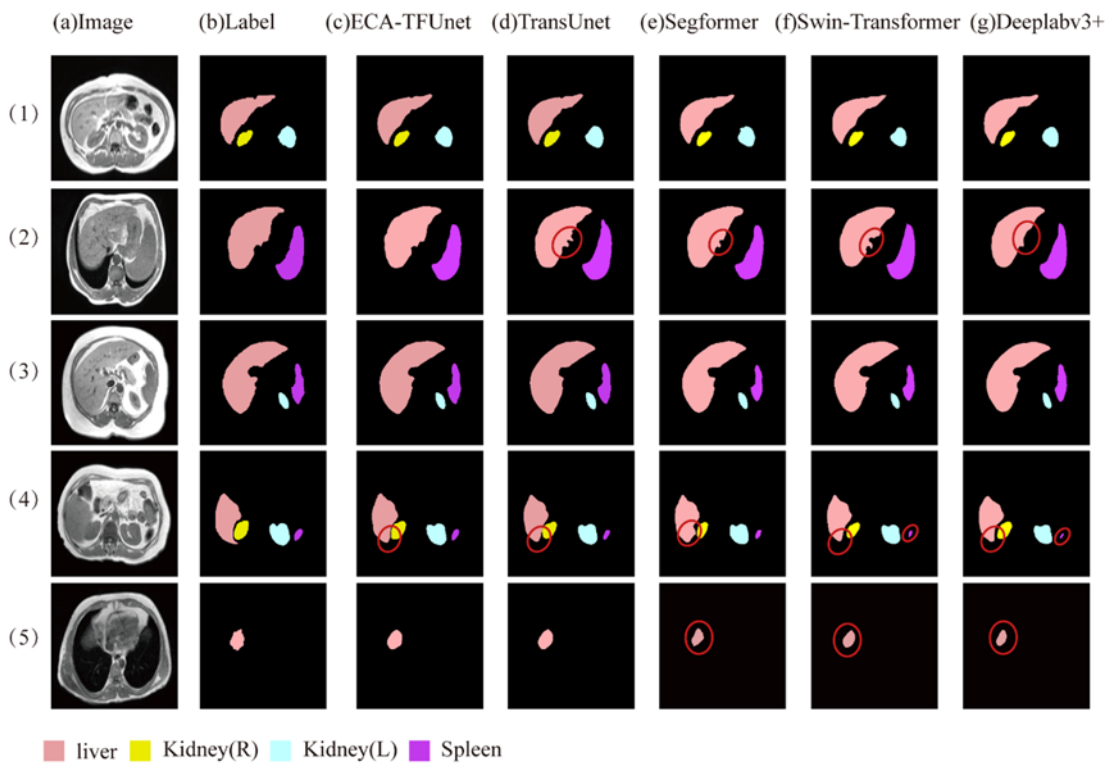

Figure 10. The segmentation visualization results on the CHAOS dataset for the top 5 models. The images in rows (1)–(5) were randomly sampled from the CHAOS dataset (including 4 different organs). column (a) is the original image, column (b) is the label image, column (c) is the segmentation result of the ECA-TFUnet and columns (d)–(g) are the results of segmentation by TransUnet, Segformer, Swin-Transformer and Deeplabv3+ models. The red circle is the place of the segmentation error.

3.6. Experiments of applying the transfer learning strategy

In transfer learning processing, data were typically divided into target and source data, with the former being directly related to the target task and the latter not. Transfer learning aims to apply the knowledge gained from the source data to the target data to improve the performance of the model on the target task. We selected the CHAOS dataset as the source data and preprocessed it using the approach described in Section 2.2 to improve transfer performance. The transfer learning strategy consisted of two stages. In the initial stage, the ECA-TFUnet model underwent pretraining on the CHAOS dataset to acquire rich general features from medical images and then provided better initialization weights for model training. In the second stage, we performed fine-tuning on anatomical sectional images of canines and reconstructed the model's segmentation head. During fine-tuning, we loaded all weight parameters obtained from the initial stage except for the segmentation head, which was initialized randomly. Figure 11 shows the Val_Loss curves of ECA-TFUnet with and without the transfer learning strategy. It can be seen that the Val_Loss curve of the model without transfer learning decreases from 0.50, while the model with transfer learning decreases from 0.25 and converges to a stable state much faster. Table 6 shows results which can be concluded that the IoU value of the model with the transfer learning is 0.41% higher than the other one.

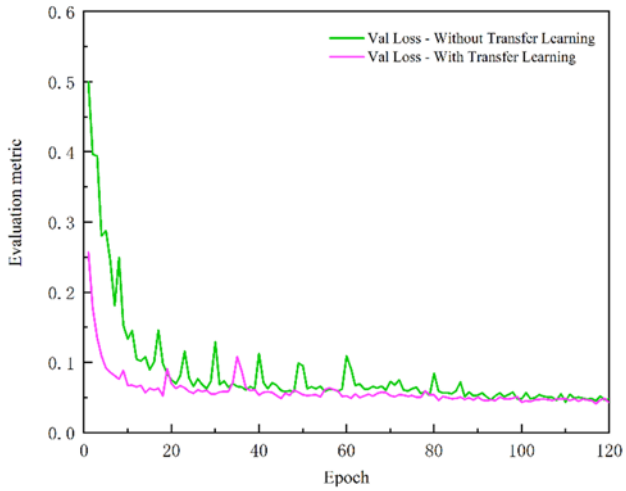

Figure 11. The Val_Loss curves of ECA-TFUnet with and without the transfer learning strategy.

Table 6. The results of ECA-TFUnet with and without the transfer learning strategy.

ECA-TFUnet model	IoU (avg)%	DSC (avg)%	ACC (avg)%
Without the transfer learning strategy	92.63	96.07	96.39
With transfer learning strategy	93.04	96.30	96.82

4. Discussion

Accurate organ segmentation in anatomical sectional images of canines enables doctors to quantitatively assess organ morphology and structural characteristics, facilitating a better understanding of canine anatomy. This plays a vital role in disease diagnosis and surgical planning.

In this study, the ECA-TFUnet model was proposed for segmenting anatomical sectional images of canines. These images contain numerous complex soft tissue structures, such as muscles and blood vessels, which have complex spatial relationships and interdependencies, and this may pose challenges to organ segmentation. To solve this problem, organ segmentation methods require strong contextual information modeling abilities to better understand the differences between organs and tissues in the image, so as to improve accuracy and reliability. The ECA-TFUnet model is a hybrid architecture that combines CNN and Transformer, incorporating the ECA block within the CNN component. Additionally, the mixed loss strategy is applied to further improve the performance of the model. Ablation experiments were conducted to validate the effectiveness and necessity of all blocks and the strategy, and the results were shown in Table 1. First, the inclusion of the Transformer block improved the IoU (avg) by 1.92%, indicating that integrating a Transformer block into the CNN can effectively enhance the segmentation accuracy. CNN models excel at capturing local detailed information, but it has weaker abilities in modeling global context [34]. In contrast, the Transformer block focuses more on global context modeling, it enables better model interaction and dependency among different local regions in the feature map [35]. Secondly, the inclusion of the ECA block resulted in a 0.6% improvement in IoU (avg). ECA block can efficiently compute feature maps and fully focus on the important channel information of the feature map, thereby it can enhance the performance of multi-organ segmentation. Thirdly, the inclusion of a mixed loss strategy resulted in an improvement of 0.48% in IoU (avg). In anatomical sectional images of canines, there is an uneven proportion distribution of

pixels of different organs, with some organs having significantly more pixels than others (e.g., liver and spinal cord). This may cause the model to be dominated by organs with more pixels, leading to poor performance in solving the class imbalance problem. Although dice loss can effectively alleviate class imbalance, it may lead to significant fluctuations in gradient updates of the prediction targets when there are partial errors in class predictions, thereby affecting training stability under specific circumstances. Cross-entropy loss can stably backpropagate gradients of different classes and effectively address the gradient vanishing problem, making the training process more stable. Thus, a mixed loss strategy combining cross-entropy loss and dice loss was applied in ECA-TFUnet, which fully combined the advantages of these two loss functions.

We demonstrated the improvement of IoU for different classes with the mixed loss strategy, as shown in Table 7. The results show that the metrics of the smaller organ class have obvious improvements, e.g., septum increases by 2.14, gallbladder by 0.72 and spinal cord by 0.66. The metrics of the larger organ class have relatively small increases, e.g., liver with a minor increase of 0.04. This suggests that the mixed loss strategy can improve the segmentation accuracy of the smaller organ class while maintaining the accuracy stability of the larger organ class, thus effectively alleviating the challenges posed by class imbalance.

Table 7. The improvement of IoU for different classes with the mixed loss strategy.

Class	Stomach	Liver	Gallbladder	Lung	Spinal Cord	Septum	Kidney (L)	Kidney (R)	Pancreas	Spleen	Intestine
Without Mixing loss Strategy	96.55	98.02	89.69	93.27	85.45	75.20	96.86	95.32	88.97	94.28	93.57
With Mixing loss Strategy	96.59	98.06	90.41	93.74	86.11	77.34	97.16	95.96	89.41	94.43	93.70
Improvement	+0.04	+0.04	+0.72	+0.47	+0.66	+2.14	+0.30	+0.64	+0.44	+0.15	+0.13

We applied the ECA-TFUnet model to the task of organ segmentation in anatomical sectional images of canines and compared it with 11 state-of-the-art models. Among them, ECA-TFUnet, TransUnet [21], Segformer [33], Swin-Transformer [32] and Deeplabv3+ [12] achieve higher scores. Figure 9 can be observed that the segmentation results of these 5 methods are all satisfactory in the images of rows (1)–(3). In the segmentation of the septum in row (4), compared with the other three models, the ECA-TFUnet and TransUnet exhibit more continuity in segmenting the septum and perform better on edge details. This may be attributed to their incorporation of multi-scale feature fusion mechanisms, which enables better handling of edge details. In the segmentation of the gallbladder in row (5), Swin-Transformer experienced a situation where the target object was not detected. TransUnet, Segformer and DeepLabv3+ all exhibited insufficient segmentation in gallbladder regions. Overall, ECA-TFUnet demonstrates enhanced capability in capturing subtle features, producing results closer to the label. The results in Table 2 indicate that our method achieves the best values in terms of IoU (avg), DSC (avg) and ACC (avg), demonstrating the effectiveness and superiority of ECA-TFUnet.

To get a more up-to-date evaluation of model performance, we conducted experiments on two datasets using the two more recent state-of-the-art models, namely SegNeXt [36] and BEiT [37]. For anatomical sectional images of canines, the IoU values of SegNeXt and BEiT are 90.22 (2.41 lower than ECA-TFUnet) and 90.04 (2.59 lower than ECA-TFUnet), respectively. For CHAOS dataset, the

IoU values of SegNeXt and BEiT are 87.22 (0.71 lower than ECA-TFUnet) and 85.36 (2.57 lower than ECA-TFUnet), respectively. Experimental results show that ECA-TFUnet achieves higher IoU values on both datasets compared to SegNeXt and BEiT. This reaffirms the excellent performance of ECA-TFUnet in organ segmentation tasks.

In Section 3.6, we attempted to further improve the performance of the ECA-TFUnet model by employing a transfer learning strategy. This strategy transfers the knowledge learned from pre-training on the source data to the target task, reducing the problem of insufficient model training caused by a lack of target data. The selection of the CHAOS dataset as the source data for transfer learning can be attributed to two primary reasons. First, the ECA-TFUnet model demonstrates exceptional performance on the CHAOS dataset, as shown in Table 4. Second, the CHAOS dataset shares a similar feature space with anatomical sectional images of canines. Figure 11 shows that ECA-TFUnet model with a transfer learning strategy achieves a lower initial loss value and faster convergence to a stable state. This reflects that the transfer learning strategy provides good initialization parameters for the model and makes the training process more efficient. Furthermore, as shown in Table 6, the addition of the transfer learning strategy resulted in the IoU (avg) improved by 0.41%, indicating that this strategy can effectively enhance the segmentation accuracy.

Although the ECA-TFUnet model can achieve impressive results, the Transformer block introduces numerous parameters, which might cause slow convergence speed. We utilize a transfer learning strategy to accelerate overall convergence, but it is unable to reduce the number of parameters. In our upcoming work, we will employ model compression techniques to reduce redundant parameters in the ECA-TFUnet.

5. Conclusions

For automated and accurate segmentation of anatomical sectional images of canines, we propose a novel ECA-TFUnet model that has advantages in both the network structure and optimization strategy. Specifically, the Transformer block can enhance the interaction and dependency of different local regions in the feature map, improving the model's representation ability. The ECA block can enhance the expression of more important channel information, improving the robustness of the model. The mixed loss strategy can alleviate the problem of class imbalance. Experiments show that the ECA-TFUnet model achieves superior segmentation performance on anatomical sectional images of canines, with IoU (avg), DSC (avg) and ACC (avg) of 92.63%, 96.07% and 96.39%, respectively, which outperforming 11 state-of-the-art models. Furthermore, the CHAOS dataset was chosen to comprehensively evaluate the segmentation performance of the ECA-TFUnet model. The results of IoU (avg), DSC (avg) and ACC (avg) reached 87.93%, 93.46% and 94.78%, respectively, which are higher than the other 11 models. These experimental results further validated the effectiveness and superiority of ECA-TFUnet. Finally, the transfer learning strategy was incorporated into ECA-TFUnet with the CHAOS dataset as source data, and IoU was improved to 93.04%. The ECA-TFUnet model enables the automatic and accurate segmentation of organs in anatomical sectional images of canines. This model provides veterinarians with accurate organ segmentation results, contributing to the efficiency of disease diagnosis. Moreover, it has potential applications in medical education that can help students quickly understand anatomical structures. Additionally, it is a critical step in the 3D reconstruction, helping to enable more complex anatomical visualization and analysis. The code for this research is available in GitHub repository: https://github.com/btbtnb/ECA-TFUnet.

Use of AI tools declaration

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

Conflict of interest

The authors declare there is no conflict of interest.

References

1. K. Karasawa, M. Oda, T. Kitasaka, K. Misawa, M. Fujiwara, C. W. Chu, et al., Multi-atlas pancreas segmentation: Atlas selection based on vessel structure, *Med. Image Anal.*, **39** (2017), 18–28. https://doi.org/10.1016/j.media.2017.03.006
2. P. F. Li, P. Liu, C. L. Chen, H. Duan, W. J. Qiao, O. H. Ognami, The 3D reconstructions of female pelvic autonomic nerves and their related organs based on MRI: a first step towards neuronavigation during nerve-sparing radical hysterectomy, *Eur. Radiol.*, **28** (2018), 4561–4569. https://doi.org/10.1007/s00330-018-5453-8
3. H. S. Park, D. S. Shin, D. H. Cho, Y. W. Jung, J. S. Park, Improved sectioned images and surface models of the whole dog body, *Ann. Anat.*, **196** (2014), 352–359. https://doi.org/10.1016/j.aanat.2014.05.036
4. J. S. Park, Y. W. Jung, Software for browsing sectioned images of a dog body and generating a 3D model, *Anat. Rec.*, **299** (2016), 81–87. https://doi.org/10.1002/ar.23200
5. K. Czeibert, G. Baksa, A. Grimm, S. A. Nagy, E. Kubinyi, Ö. Petneházy, MRI, CT and high resolution macro-anatomical images with cryosectioning of a Beagle brain: creating the base of a multimodal imaging atlas, *PLoS One*, **14** (2019), e0213458. https://doi.org/10.1371/journal.pone.0213458
6. X. Shu, Y. Y. Yang, B. Y. Wu, A neighbor level set framework minimized with the split Bregman method for medical image segmentation, *Signal Process.*, **189** (2021), 108293. https://doi.org/10.1016/j.sigpro.2021.108293
7. X. Shu, Y. Y. Yang, J. Liu, X. J. Chang, B. Y. Wu, ALVLS: Adaptive local variances-Based levelset framework for medical images segmentation, *Pattern Recogn.*, **136** (2023), 109257. https://doi.org/10.1016/j.patcog.2022.109257
8. S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, et al., A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises, *Proc. IEEE*, **109** (2021), 820–838. https://doi.org/10.1109/JPROC.2021.3054390
9. A. Majumdar, L. Brattain, B. Telfer, C. Farris, J. Scalera, Detecting intracranial hemorrhage with deep learning, in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, (2018), 583–587. https://doi.org/10.1109/EMBC.2018.8512336
10. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

11. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 4700–4708. https://doi.org/10.1109/CVPR.2017.243
12. L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 801–818.
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*, Springer, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
14. D. Schmid, V. B. Scholz, P. R. Kircher, I. E. Lautenschlaeger, Employing deep convolutional neural networks for segmenting the medial retropharyngeal lymph nodes in CT studies of dogs, *Vet. Radiol. Ultrasound*, **63** (2022), 763–770. https://doi.org/10.1111/vru.13132
15. J. Park, B. Choi, J. Ko, J. Chun, I. Park, J. Lee, et al., Deep-learning-based automatic segmentation of head and neck organs for radiation therapy in dogs, *Front. Vet. Sci.*, **8** (2021), 721612. https://doi.org/10.3389/fvets.2021.721612
16. 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*, (2021), 205–218. https://doi.org/10.1007/978-3-031-25066-8_9
17. Y. Xu, X. He, G. Xu, G. Qi, K. Yu, L. Yin, et al., A medical image segmentation method based on multi-dimensional statistical features, *Front. Neurosci.*, **16** (2022), 1009581. https://doi.org/10.3389/fnins.2022.1009581
18. 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, preprint, arXiv:2010.11929.
19. 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
20. S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 6881–6890. https://doi.org/10.1109/CVPR46437.2021.00681
21. 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.
22. B. Li, S. Liu, F. Wu, G. Li, M. Zhong, X. Guan, RT-Unet: An advanced network based on residual network and transformer for medical image segmentation, *Int. J. Intell. Syst.*, **37** (2022), 8565–8582. https://doi.org/10.1002/int.22956
23. H. Wang, P. Cao, J. Wang, O. R. Zaiane, Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **36** (2022), 2441–2449. https://doi.org/10.1609/aaai.v36i3.20144
24. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 11534–11542. https://doi.org/10.1109/CVPR42600.2020.01155
25. A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P. H. Conze, V. Groza, et al., CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation, *Med. Image Anal.*, **69** (2021), 101950. https://doi.org/10.1016/j.media.2020.101950

26. 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.
27. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems*, **30** (2017).
28. H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 2881–2890. https://doi.org/10.1109/CVPR.2017.660
29. J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, et al., Dual attention network for scene segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 3146–3154.
30. Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. https://doi.org/10.1109/ICCVW.2019.00246
31. Y. Yuan, X. Chen, J. Wang, Object-contextual representations for semantic segmentation, in *European Conference on Computer Vision*, Springer, (2020), 173–190. https://doi.org/10.1007/978-3-030-58539-6_11
32. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986
33. E. Z. Xie, W. H. Wang, Z. D. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, in *Advances in Neural Information Processing Systems*, **34** (2021), 12077–12090.
34. M. D. Alahmadi, Medical image segmentation with learning semantic and global contextual representation, *Diagnostics*, **12** (2022), 1548. https://doi.org/10.3390/diagnostics12071548
35. J. Fang, C. Yang, Y. Shi, N. Wang, Y. Zhao, External attention based TransUNet and label expansion strategy for crack detection, *IEEE Trans. Intell. Transp. Syst.*, **23** (2022), 19054–19063. https://doi.org/10.1109/TITS.2022.3154407
36. M. H. Guo, C. Z. Lu, Q. Hou, Z. Liu, M. M. Cheng, S. M. Hu, SegNeXt: Rethinking convolutional attention design for semantic segmentation, in *Advances in Neural Information Processing Systems*, **35** (2022), 1140–1156.
37. H. Bao, L. Dong, S. Piao, F. Wei, BEiT: BERT pre-training of image transformers, preprint, arXiv:2106.08254.

AIMS Press