



SAA-SDM: Neural Networks Faster Learned to Segment Organ Images

Chao Gao^{1,2} · Yongtao Shi^{1,2} · Shuai Yang^{1,2} · Bangjun Lei^{1,2}

Received: 5 August 2023 / Revised: 16 October 2023 / Accepted: 18 October 2023 / Published online: 10 January 2024
© The Author(s) under exclusive licence to Society for Imaging Informatics in Medicine 2024

Abstract

In the field of medicine, rapidly and accurately segmenting organs in medical images is a crucial application of computer technology. This paper introduces a feature map module, Strength Attention Area Signed Distance Map (SAA-SDM), based on the principal component analysis (PCA) principle. The module is designed to accelerate neural networks' convergence speed in rapidly achieving high precision. SAA-SDM provides the neural network with confidence information regarding the target and background, similar to the signed distance map (SDM), thereby enhancing the network's understanding of semantic information related to the target. Furthermore, this paper presents a training scheme tailored for the module, aiming to achieve finer segmentation and improved generalization performance. Validation of our approach is carried out using TRUS and chest X-ray datasets. Experimental results demonstrate that our method significantly enhances neural networks' convergence speed and precision. For instance, the convergence speed of UNet and UNET++ is improved by more than 30%. Moreover, Segformer achieves an increase of over 6% and 3% in mIoU (mean Intersection over Union) on two test datasets without requiring pre-trained parameters. Our approach reduces the time and resource costs associated with training neural networks for organ segmentation tasks while effectively guiding the network to achieve meaningful learning even without pre-trained parameters.

Keywords Medical image segmentation · Semantic segmentation · Neural network · Accelerating neural network learning

Introduction

In medical image segmentation, the current mainstream method uses a neural network (NN) for semantic segmentation [1, 2]. Compared with traditional segmentation methods, such as manually designed boundary operators [3, 4], NN can automatically learn features through a large amount of data to perform related tasks more accurately [5].

Secondly, compared with some level-set methods that need to embed prior information [6, 7] or high-level operations [8], NN can also deal with more complex and abstract features. It can learn more advanced feature representations and have strong generalization abilities [2]. In addition, neural networks can perform related tasks on different dimensions of

features by using convolutional layers and other types of hierarchical structures, enabling neural networks to capture the local and global features of images, thereby improving the accuracy of segmentation [9, 10].

The application of NN in medical image segmentation has many advantages. However, due to its width and depth scale, the amount of matrix calculation and numerical length required is tremendous. Hence, the training of NN usually requires large-scale computing power to support. In recent years, several works have proposed methods and techniques to speed up the training of, or predictions by, neural networks from various perspectives [11, 12].

Compared with the various photos in our daily lives, although the style of medical images is single and different from optical imaging, the imaging methods used in medical images often have significant anatomical noise, global noise, and fuzzy boundaries [2]. Although this property can slow down the neural network's learning speed, the NN's learning process can be simplified because human organs tend to have relatively fixed shapes and positions. Islam et al. [13] explored how neural networks contain positional and semantic information and found that many

✉ Yongtao Shi
ytshi@ctgu.edu.cn

¹ College of Computer and Information Technology, China Three Gorges University, Yichang Hubei 443002, China

² Hubei Key Laboratory of Intelligent Vision Monitoring for Hydropower Engineering, China Three Gorges University, Yichang Hubei 443002, China

irrelevant training neural networks would give semantic feedback to a single color gradient image. That is an exciting experiment and conclusion, which shows the powerful learning ability of neural networks. Therefore, there should be a more straightforward way to speed up the learning of semantic information that neural networks spend a lot of time learning.

Unlike previous studies, this paper proposes a novel neural network convergence acceleration method for medical image segmentation. The main idea is to embed geometric representations into the network training, which can significantly reduce the amount of data so that the neural network can pay attention to the target as soon as possible to accelerate the training process. Specifically, we propose the Strength Attention Area Signed Distance Map (SAA-SDM), which extracts semantic features from the data and adds them to the training of the neural network to assist the training. Our experiments show that this behavior changes the neural network's original learning and prediction strategy, which makes the network learn from the initial disordered global search type to the general to individual learning. In addition, to maintain the generalization performance of the network and ensure the stability and efficiency of the network training process, we gradually weaken SAA-SDM during the training process. Subsequently, the neural network learns features from the image and de-pends on this module gradually.

We apply this approach to several popular neural network frameworks and validate it using two medical image datasets, X-ray and ultrasound images. The medical imaging principles of these two datasets are widely used and representative of the medical field. For instance, ultrasound images exhibit broader applicability and reproducibility but face higher noise levels and lower contrast challenges. Experiments show that our method produces good results in almost all tests. Moreover, our module has better guided the transformer architecture network to train without pre-training parameters and has higher convergence speed and pixel classification accuracy. Therefore, the main contributions of this paper can be summarized as follows:

- SAA-SDM module is proposed and added to the training, which can make the neural network converge to a high accuracy level faster.
- This type of training guides the network to learn a new policy. Experiments show that this learning method is more efficient.
- This method does not conflict with other current speedup schemes and does not incur more resource consumption, so more policies can be considered simultaneously to improve training efficiency.

Materials and Methods

Overview

Our training and inference process is shown in Fig. 1, which refines the semantic information of the mask to reduce the data scale. Then, multiple components of the point set were extracted and expanded to form SAA-SDM. Finally, it was introduced into the network training according to the training strategy. Although Liu et al. [14] found that using explicit sequential position encodings can improve regression performance in neural networks, they did not propose more encodings to aid other tasks. Previously, some work added auxiliary task heads at the end of the network to generate more significant gradients that could be learned. For example, Li et al. [15] and Lui et al. [16] add SDM at the end of the network to strengthen supervision of the NN learning to achieve accuracy improvement. Therefore, this is a way that the NN can understand the encoding way. This approach provides encoding feasibility for the segmentation task compared to the pure coordinate regression task.

In this section, the theory of SAA-SDM and its composition is introduced in “SAA-SDM Module”. “Training Process” introduces the training process of SAA-SDM after joining the network and its exit. “Metrics” will introduce multiple accuracy and training acceleration indicators for evaluating the model to measure network prediction accuracy from multiple perspectives.

SAA-SDM Module

The main task of the proposed SAA-SDM module is to help the network build a faster and better initial cognition. First, contour extraction sampling from the masks of the training data is required. Let the sampling results of all images be $\mathbf{Y}_{lm} = \{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_s\}$ and s be the number of samples. For the i th sample, there is $\mathbf{Y}_{lm}^i = \{p_1, p_2, \dots, p_n\}$, where n is the number of sampling points. Since the size and shape of the target in each image are different, in order to reduce the learning difficulty of the neural network caused by the numerical difference, it is necessary to perform the exact translation and normalization operation on all coordinates in \mathbf{Y}_{lm}^i by referring to the relative translation of its image center at the origin. As shown in the following equation, for any sample, we have the following:

$$\mathbf{Y}_{c_lm}^i = \mathbf{Y}_{lm}^i - [\min(\mathbf{Y}_{lm}^i) + \max(\mathbf{Y}_{lm}^i)]/2 \quad (1)$$

$$\mathbf{Y}_{n_lm}^i = \frac{\mathbf{Y}_{c_lm}^i - \min(\mathbf{Y}_{c_lm}^i)}{\max(\mathbf{Y}_{c_lm}^i) - \min(\mathbf{Y}_{c_lm}^i)} \quad (2)$$

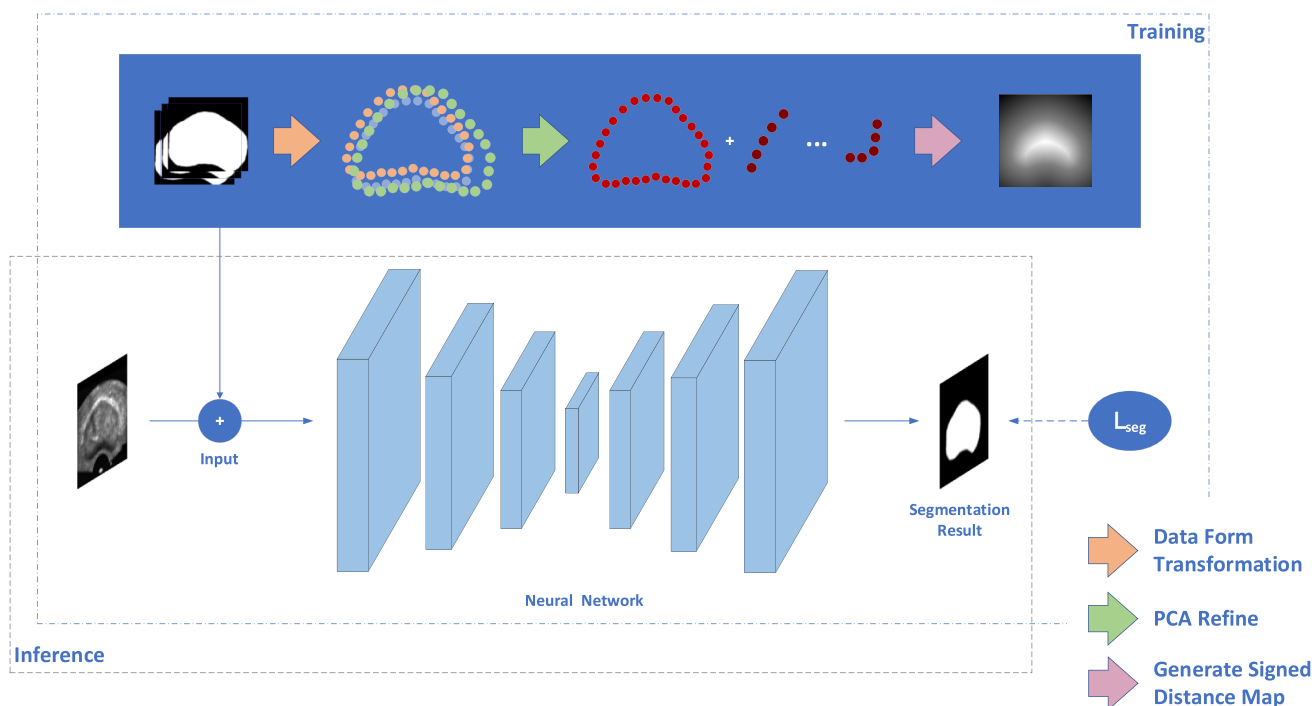


Fig. 1 Schematic of the framework for training and inference. During training, the strength attention area signed distance map (SAA-SDM) generated using the PCA method is input into the network alongside

the original image. During the inference stage, the neural network remains in its original state

is given below to form a new data point set $\mathbf{Y}_{n_lm} = \{ \mathbf{Y}_{n_lm}^1, \mathbf{Y}_{n_lm}^2, \dots, \mathbf{Y}_{n_lm}^s \}$. In order to find the orthogonal feature with the most significant variance under each feature point, the covariance matrix Σ of all the given training samples needs to be calculated. The mean value of the corresponding points of each sample can be obtained from Eq. 3, and then the covariance matrix Σ of all samples can be obtained from Eq. 4. Its ordinal eigenvalue set λ , and the corresponding eigenvector set \mathbf{p} can be obtained.

$$\tilde{\mathbf{Y}} = \frac{1}{s} \sum_{i=1}^s \mathbf{Y}_{n_lm}^i \tag{3}$$

$$\Sigma = \frac{1}{s} \sum_{i=1}^s (\mathbf{Y}_{n_lm}^i - \tilde{\mathbf{Y}})(\mathbf{Y}_{n_lm}^i - \tilde{\mathbf{Y}})^T \tag{4}$$

The symbol T represents the transpose operation of the matrix. Since the reference value of the orthogonal feature with less influence is relatively small and the organ size usually follows a normal distribution, when generating SDM, the value of the orthogonal feature is usually normalized from 0 to 1, and the difference between adjacent pixel values is reduced to less than 0.005. The pixel value after 0–1 normalization is at least 0.004. The reason makes feature gaps outside the 90% interval challenging to detect. Therefore, the first eigenvalues are limited enough

to represent most of the data, that is, to satisfy the following equation:

$$\arg \min \frac{\sum_{j=1}^m \lambda_j}{\sum_{i=1}^n \lambda_i} \geq 90\% \tag{5}$$

According to the principle of PCA, the coefficients on each feature vector used to describe the original sample can be found here. In other words, for m feature vectors, m coefficients $\alpha_i = [\alpha_1^i, \alpha_2^i, \dots, \alpha_m^i]^T$ can be found so that the corresponding sample $\mathbf{Y}_{n_lm}^i = \alpha_i^T \cdot \mathbf{p}$ can be restored as much as possible, where i is the sequence number of the sample. All the coefficients can form a matrix $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_s]$ together, and then the approximate representation of the contour of all the samples can be obtained:

$$\mathbf{Contour} = \alpha^T \cdot [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_s] \tag{6}$$

Through Eq. 6, two vectors, \mathbf{c}_{\max} and \mathbf{c}_{\min} , are formed by the largest and smallest absolute values of **Contour** according to row, respectively, which means that the maximum area of the target in all samples is no more than the area enclosed by the \mathbf{c}_{\max} point set, and the minimum area is no less than the area enclosed by the \mathbf{c}_{\min} point set; that is, the target to be predicted can be expected, and its

boundary point set \mathbf{bd} satisfies the conditions of Eqs. (7) and (8), and it is combined into a set S :

$$c_{\min}^i \in \mathbf{c}_{\min} \leq \mathbf{bd}^i \leq c_{\max}^i \in \mathbf{c}_{\max} \quad (7)$$

$$\text{area}(\mathbf{c}_{\min}) \leq \text{area}(\mathbf{bd}) \leq \text{area}(\mathbf{c}_{\max}) \quad (8)$$

Now, we define SAA-SDM module, which needs to define an extended level set function (LSF). $\phi : \Omega \rightarrow R$ can be defined as the signed distance map (SDM), but the difference is that we do not define an exact SDM, but rather an interval of interest:

$$\phi(x) = \begin{cases} -\inf_{a \in \partial S} \|b - a\|_2, & b \in S_{in} \\ 0, & b \in \partial S \\ \inf_{a \in \partial S} \|b - a\|_2, & b \in S_{out} \end{cases} \quad (9)$$

where a is another pixel in the segmentation mask, ∂S represents for the area where an object's boundary may exist, not its exact boundary. In the classic SDM, the zero SDM value of a point means that the point is on the surface of the target object. But we expand this part even larger, and almost the boundaries of what is possible are included in it. S_{in} and S_{out} denote the non-boundary inner and outer regions of the target object. Since input images have various fields of view and organ volumes, we further normalize the $\phi(x)$ of each pixel to the range $[-1, 1]$.

Training Process

In theory, adding the SAA-SDM module to the input end of a neural network weakens the ratio of the original image throughout the input, requiring convolutional operations to consider information from both the SAA-SDM module and the original image pixels. Here, X_{in} represents the original input image, and the neural network F can be viewed as a collection of multiple nonlinear functions $\{f_1, f_2, \dots, f_n\}$. Typically, nonlinear functions are defined as a nesting of one or more linear functions $l(g)$, regularization functions $norm(g)$, and nonlinear units $act(g)$, as shown in Eq. 10:

$$f(x; \phi(x)) = act(norm(l(x; \phi(x)))) \quad (10)$$

Introducing SAA-SDM results in a change of sign in the target region of the input. Using the method introduced in "SAA-SDM Module", it can be observed that this implicitly encodes positional information about the target. Equation 9 illustrates this phenomenon, where the absolute value of $|\phi(x)|$ indicates the probability of whether or not it is a particular target. Therefore, in the initial stages of network training, its output results can be approximated and expressed as Eq. 11:

$$F(\phi(x)) \sim F(x; \phi(x)) \quad (11)$$

The formula means that the neural network can distinguish between the interior and exterior regions of the target simply by evaluating the signs of $\phi(x)$. Compared to the original training approach, including SAA-SDM significantly reduces the difficulty for the neural network to learn the semantic information of the target. In the former case, semantic relationships between pixels could only be determined based on vague pixel relationships.

Nevertheless, adding the SAA-SDM module to the network for training is more than just a one-and-done deal. Although convolutions dynamically assign weights to each element, stable networks always accept that SAA-SDM can have harmful effects. For example, the initial network is like a toddler, requiring support to walk upright; an adult using the toddler's tools would be a barrier. The SAA-SDM is not specific to the data input, so the network needs to realize that the segmentation must be based on the original information and nothing else. However, this is challenging to do during training.

For a more careful explanation, let us simplify the problem. As shown in Fig. 2, the blue block is the input, and the SAA-SDM is the purple block. The gray and white blocks represent the convolution parameters. By setting padding = 1, we can keep the input and output sizes the same. In the initial learning stage ①, SAA-SDM is the primary basis for network prediction because the first few rounds of network training are insufficient to form a good cognition of the original images compared with SAA-SDM. In the second stage of learning ②, the uncontroversial parts of the prediction will no longer have disagreements, so the neural network backpropagation will not produce large gradients. The controversial areas will make the neural network more significantly impact the assigned weights. The weight parameters of the network will gradually be weighted towards the blue patches. We expect the network in stage ③ to realize that the loss reduction is focused on distinguishing the values in the blue patches and ignoring the purple patches, which are not directly related to the input, where the problem arises. The blue patch represents all the pixel information of the original image. However, in medical images, due to noise and artifacts, the neural network cannot entirely rely on the pixel information for prediction, leading to the neural network's misjudgment of the relationship between the two. The purple patch still has usable information, which cannot be discarded entirely. It is precisely because of this misjudgment that leads to the situation that the network accuracy may be reduced after adding SAA-SDM. The solution is that our training process is designed to let SAA-SDM phase-like exit scheme.

Dynamically changing the input used for training may harm the stability of the network training. For example, if the original data is directly used after a certain time point, the neural network will collapse due to loss of reference. A smooth exit plan is, therefore, essential. The input X of the

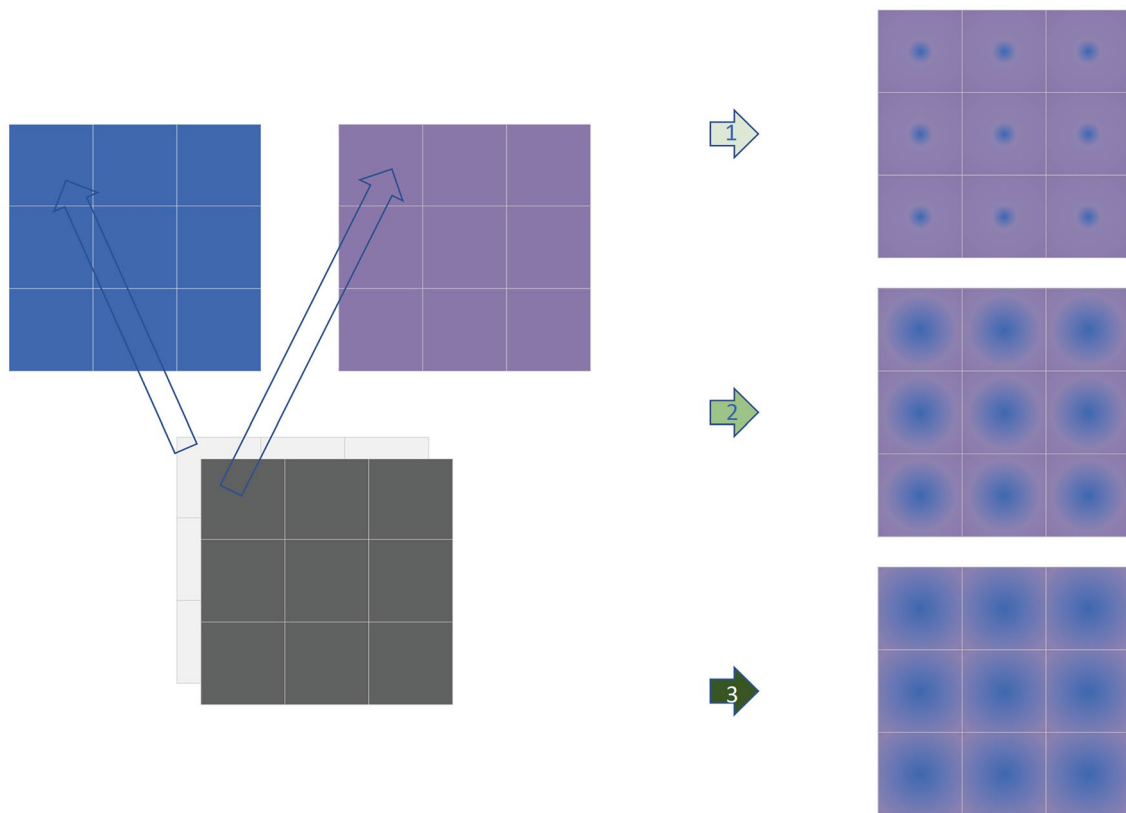


Fig. 2 Schematic representation of the weights of the convolution parameters. The more colors a given color has, the more it is represented in the feature map

neural network consists of two parts, as shown in the following equation:

$$X = X_{in} \oplus X_{SAA} \cdot \varepsilon \tag{12}$$

where ε is the weight coefficient of X_{SAA} , which is used to control the participation of X_{SAA} in the total input X . When $\varepsilon = 0$, the input of the neural network is returned to the original input. The deterministic annealing process, by which *decay_rate* is slowly increased, is expected to help the optimization process to avoid poor local minima [17]. This scheme shows good results in pseudo-label participation training [18]. Nevertheless, its setting is a static phase of adjustment. In order to be closer to the actual situation of network training, we dynamically give a dynamic design according to the feedback of the network:

$$\varepsilon = \begin{cases} \varepsilon & , \quad \text{if } \frac{|val_acc_{t-1} - val_acc_t|}{val_acc_t} \geq 0.05 \\ \varepsilon - decay_rate & , \quad \text{elseif } decay_thresh \geq decay_rate \\ 0 & , \quad \text{otherwise} \end{cases} \tag{13}$$

Among them, the value of ε is initialized to 1. When the prediction accuracy gap on the test dataset exceeds 5% for

two consecutive times, it is considered unstable. It must keep the original value to continue training until the network output is stable. *decay_thresh* is a threshold. After the decay rate is reduced to the threshold, the neural network will no longer react to the SAA-SDM in the input. At this time, setting SAA-SDM to 0 does not affect the discrimination of the network to obtain a good training process.

Metrics

We used commonly used evaluation metrics, and our own provided an evaluation metric to measure the method’s superiority. Commonly used evaluation metrics include mean dice similarity coefficient (mDSC, also called mDice), mean Intersection over Union (mIoU), accuracy, precision, recall, and kappa.

The prediction of the neural network was compared with the True value to obtain four indicators: true positive (TP), true negative (TN), false positive (FP), false negative (FN), then.

1. Dice:

$$mDice = \frac{1}{s + 1} \sum_{i=0}^s \frac{2 \times TP}{2 \times TP + FN + FP} \tag{14}$$

2. IoU:

$$mIoU = \frac{1}{s+1} \sum_{i=0}^s \frac{TP}{TP+FN+FP} \quad (15)$$

3. Accuracy:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (16)$$

4. Precision and recall:

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN} \quad (17)$$

5. Kappa:

$$kappa = \frac{Accuracy - Recall}{1 - Recall} \quad (18)$$

Since the prediction accuracy of the neural network on the test set fluctuates, if the fluctuation range of the neural network is set to be less than 1% in a long training time, the neural network training is considered to have reached the convergence state. The round that enters this range for the first time is recorded as e_1 , and the mIoU of the test set of this round is p_1 . If e_2 is recorded in another network training, which should meet the mIoU $p_1 - 1\% \leq p_2 \leq p_1 + 1\%$ for the first time, then the following can be calculated:

$$Acceleration = \left(1 - \frac{e_1}{e_2}\right) \times 100\% \quad (19)$$

For example, if a neural network converges at 120 rounds in state one and the mIoU is 0.9, then the mIoU should not exceed 0.91 on all test sets of the network. When the network is in state two, and the accuracy is greater than 89% for the first time at round 160, it is considered that the convergence round is 160; that is, $e_1 = 120$, $e_2 = 160$, then

$$Acceleration = \left(1 - \frac{e_1}{e_2}\right) \times 100\% = \left(1 - \frac{120}{160}\right) \times 100\% = 25\%.$$

Statistical Testing Methods

The training process of neural networks typically involves random weight initialization, random selection of sample order, and regularization terms, among other stochastic factors. This results in neural networks producing outcomes and metrics that do not follow a precise data distribution. Furthermore, despite the ability to conduct independent repeated experiments, neural networks often require lengthy training, resulting in limited neural network samples available for statistical analysis. Approximating all data to a normal distribution through the law of large numbers becomes challenging.

To thoroughly investigate the role of the proposed module in the network, we chose to use the Brown-Forsythe Test and Mann-Whitney U test in combination to assess the module's effectiveness within the network. Compared to other hypothesis testing methods, the Brown-Forsythe test and Mann-Whitney U test are non-parametric statistical tests, making no assumptions about the data distribution. These test methods make them compelling even when the data does not adhere to assumptions like normality. The Mann-Whitney U test also imposes no particular restrictions on sample size, making it suitable for small sample testing. This approach aims to evaluate the extent to which experimental conclusions are acceptable.

We collected six samples for each experimental group, and all parameters except for the control module were set to fixed and identical values in each experiment. The hypotheses of the two test methods are shown in Table 1. Furthermore, the p -value < 0.05 was considered statistically significant.

The H hypothesis is shown in the table above. Because it is necessary to verify the growth rate ratio of the control samples, confirm that the fluctuation in convergence rounds between the two has a reasonable range of variance, and eliminate bias in conclusions due to variance, the Brown-Forsythe Test needs to be conducted on both sets of data before performing the Mann-Whitney U test. The detailed process for the test is as follows:

Table 1 Meaning of null hypothesis in hypothesis testing

Test methods		Means
Brown-Forsythe test	H0	The sample variances are equal across different groups
	H1	The sample variances are not equal across different groups
Mann-Whitney U test	H0	Two independent sample groups come from the same distribution, and there is no significant difference between the two sets of data
	H1	Two independent sample groups come from different distributions, and there is a significant difference between the two sets of data

1. Obtain two sets of observed samples, X_1 and X_2 , from neural networks with and without SAA-SDM, respectively;
2. Perform the Brown-Forsythe Test on the observed samples X_1 and X_2
 - (a) If the Brown-Forsythe Test results in $p < 0.05$, reject the null hypothesis, indicating that the samples do not yield valid conclusions;
 - (b) If the Brown-Forsythe Test results in $p > 0.05$, accept the null hypothesis and proceed to (c);
 - (c) Conduct the Mann–Whitney U test on the observed samples X_1 and X_2
 - (i) If $p > 0.05$ is obtained, indicating no significant difference between the samples, valid conclusions cannot be drawn;
 - (ii) If $p < 0.05$ is obtained, indicating a significant difference between the samples, proceed to step 3;
3. Calculate the two groups' sample means, \bar{X}_1 and \bar{X}_2 , and compute the mean ratio $r = \bar{X}_1 / \bar{X}_2$. Then, align the mean of X_2 to \bar{X}_1 , resulting in $X'_2 = r \times X_2$;
4. Conduct the Brown-Forsythe Test on samples X_1 and X'_2
 - (a) If the Brown-Forsythe test yields $p < 0.05$, reject the null hypothesis, and valid conclusions cannot be drawn;
 - (b) If the Brown-Forsythe test results in $p > 0.05$, proceed to step (c);
 - (c) Perform the Mann–Whitney U test on samples X_1 and X'_2
 - (i) If the Mann–Whitney U test results in $p < 0.05$, indicating a significant difference between the aligned data samples, valid conclusions cannot be drawn;
 - (ii) If the Mann–Whitney U test results in $p > 0.05$, suggesting that it is impossible to determine a significant difference between the aligned data samples, proceed to determine the range of differences;
 - (d) Adjust the floating range of r found in step 3 to identify its maximum and minimum values that result in both the Brown-Forsythe test and Mann–Whitney U test yielding $p > 0.05$. This range of $r \in [r_{\min}, r_{\max}]$ determines the growth rate significance level.

Experiment

Datasets Description

In order to verify the accuracy of the proposed method, we apply it to two 2D medical image datasets for testing. The network can be extended to 3D images like the work reported in [19]. As mentioned in the previous section, our approach is not limited by the form of the network framework. Next, we detail the two datasets we used for our experiments, the TRUS and Pulmonary Chest X-Ray datasets.

TRUS Datasets

The TRUS dataset [20] contains TRUS images of 108 patients, all with discernible information removed. All ultrasound images used in this dataset were acquired using the same setup using a Philips HDI 5000 SonoCT imaging system, with different patients corresponding to their images separately. The size of each image is 768×576 pixels, and the size of the pixels is $0.137 \text{ mm} \times 0.137 \text{ mm}$. To generate the training data, the expert described each Truth of the prostate image, and the Ground Truth of all pictures was annotated in a landmark manner with the number of marked points in each landmark being 100. Before the experiments, we processed this into pixel-wise segmentation labels common to neural networks and object detection box data by interpolating the pixels between any two truth marker points until we had a fully closed pixel-wise contour, then performing padding on the interior.

Pulmonary Chest X-Ray Datasets

The National Library of Medicine, Maryland, USA, created the standard digital image database for tuberculosis in collaboration with Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China. The chest X-rays are from outpatient clinics and were captured as part of the daily routine using Philips DR Digital Diagnose systems. China Set-The Shenzhen set-Chest X-ray Database [21] provides 326 normal lung images; the X-rays are provided in PNG format. Their size can vary but is approximately $3 \text{ K} \times 3 \text{ K}$ pixels. The Montgomery County X-ray Set [22] provides 80 normal lung images; the X-rays were captured with a Eureka stationary X-ray machine (CR) and are provided in portable network graphics (PNG) format as 12-bit gray level images. The size of the X-rays is either 4020×4892 or 4892×4020 pixels. All images are de-identified and available in DICOM format

and PNG format. One of the main tasks of releasing this public dataset is lung segmentation experiments.

Implementation Details

We used PaddlePaddle deep learning framework, NVIDIA Tesla V100 16G GPU; due to the limited hardware capacity available and to ensure that our actions do not cause the image to lose recognizable, influential large-scale content, we scaled China set-the Shenzhen Set-Chest X-ray database from $3\text{ K} \times 3\text{ K}$ pixels to 512×512 pixels. In addition, we crop the irrelevant image edges to 4020×4020 pixels according to the Ground Truth in the Montgomery County X-ray Set dataset and then scale them to 512×512 pixels. During the training of the neural network, we applied random data augmentation to the data, which included image brightness adjustment, contrast range adjustment, and saturation adjustment at 40% of the time, all controlled within 60% of the original image.

The relative treatment of the dataset is the same; we divide the data 9:1. That is, 90% of the data is used for training, and 10% is used for testing the training results. All networks used the same CrossEntropy Loss to ensure a fair comparison, with an initial learning rate of 0.01 and momentum of 0.9. Stochastic gradient descent was used for network optimization, and the L2 norm of the network weights (W , w , h) with a decay coefficient 0.001. In the simulated annealing algorithm, we set the initial value of ϵ to 1, and the initial value of *decay_thresh* is set to 0.6. The neural network was trained for at least 200 iterations (iters) in all experiments until accuracy convergence on the test set was observed.

Ablation Experiments and Results

In this section, we will compare the classical and mainstream segmentation convolutional neural networks UNET [23], UNet++ [24], UNET3+ [25], U2NET [26], SegNet [27], and Attention UNet [28] with the current popular Transformer-based segmentation networks SegFormer [29], Segmenter [30]. In all experiments, we embedded our proposed method into these mainstream segmentation networks and compared it with its original network framework. We must note that we do not compare network performance, as the networks determine this.

Qualitative Analysis

Due to the leadership of UNet in the industry, we first show in detail the network interrupt test prediction results on the test set every five rounds during the first 20 iters of training this networks. For easy observation, we output a heatmap of the segmentation results to determine where the network focuses on and find an intuitive inference strategy for the neural network.

The table shows the results of the predictions on the four test sets. The first column is the original image, the second column is the ground truth (GT), and the third to sixth columns are the samples on different rounds.

It can be seen from Table 2 that after our method is used, the neural network quickly focuses on the relatively correct location, which is guided by SAA-SDM and protected from the influence of the surrounding irrelevant regions. Specifically, after adding SAA-SDM, when iter is 5, the strong confidence region of the neural network is only inside the target (the second row of data in each test map in Table 2), which is different from the prediction result of the original network (the first row of data in each test map in Table 2), which first focuses on the entire TRUS image region (the sector area with information). From the network prediction results, when iter is 10, the network with SAA-SDM has a strong confidence in the target's interior. It starts to focus on the refined segmentation of the boundary region. On the contrary, the network trained originally does not show this behavior in the following multiple rounds of the output. From the evolution process shown by multiple iters in Table 2, we find that the neural network changes the original strategy after joining ours. In the processing of data, the original strategy of the neural network is to gradually shrink the target from coarse to fine and remain "suspicious" of the surrounding noise because the short training process is challenging to make the network grasp the high-level semantic information so that it cannot believe the correctness of the actual target position with a high degree of confidence. Ours explicitly provides a range. Although this does not constitute a strong range constraint, the semantic information is given by SAA-SDM and embedded, which makes the training process of the neural network transform from commonality to individuality so that the initial loss of the network is relatively small, which significantly reduces the training difficulty and dramatically improves the accuracy.

Quantitative Analysis

The quantitative results of the TRUS dataset run in the individual networks are shown in Table 3. In addition, from the perspective of the accuracy of the network trained to convergence, the accuracy of the network did not decrease significantly after joining our method. It also needs to reflect not the lack of generalization performance. However, the increase in accuracy during training was significant.

In the current prevalent framework of various U-shaped networks built based on CNN, the final accuracy of our network remained stable due to the introduction of the proposed module. On the contrary, the method proposed in this paper dramatically improves the convergence speed of the neural network under the premise of ensuring accuracy. In the observed results, using the UNet++ neural network structure to introduce the proposed

Table 2 Sample survey form for interrupt testing in training

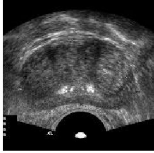

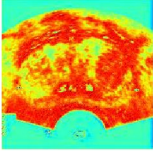
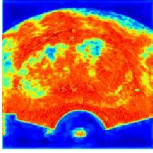
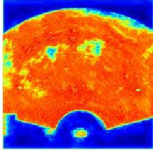
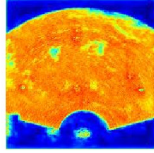
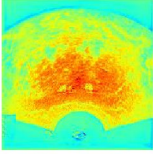
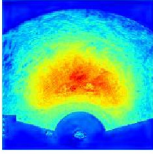
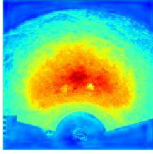
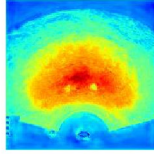
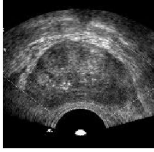

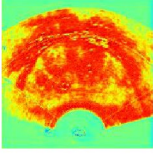
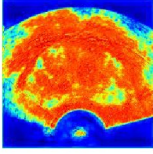
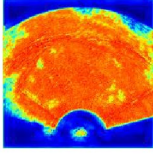
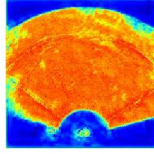
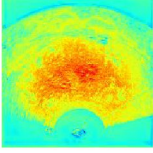
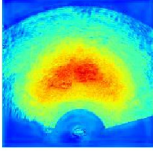
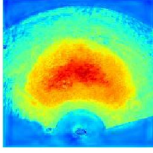
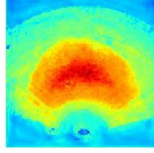


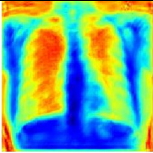
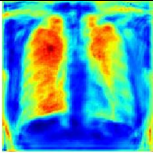
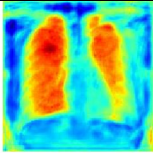
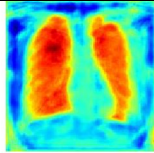
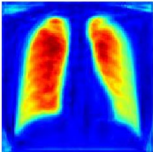

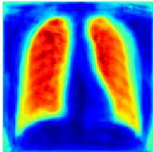
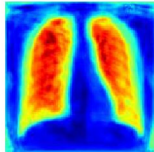


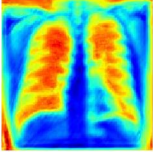
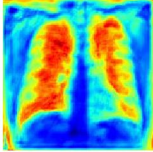
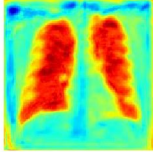
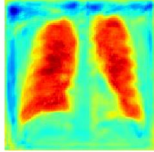
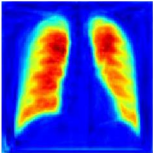


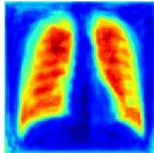
IMG	Mask	5	10	15	20
					
		UNet			
					
		UNet + SAA-SDM (ours)			
					
		UNet			
					
		UNet + SAA-SDM (ours)			
					
		UNet			
					
		UNet + SAA-SDM (ours)			
					
		UNet			
					
		UNet + SAA-SDM (ours)			

Table 3 Test result data on TRUS Datasets. Each neural network has two rows of conclusion data. The first row of data corresponds to the original neural network framework proposed by the authors. In contrast, the second row of data represents the test results after incorporating our method SAA-SDM

Name	mDice	mIoU	Accuracy	Precision	Recall	Kappa	Acceleration
UNet	0.9523	0.9089	0.9751	0.9503	0.9543	0.9355	-
	0.9493	0.9035	0.9734	0.9418	0.9573	0.9313	42.86%
SegNet	0.6499	0.4813	0.7981	0.592	0.7203	0.5099	-
	0.9272	0.8643	0.9616	0.9154	0.9393	0.9012	∞
Attention UNet	0.9413	0.91372	0.9766	0.9677	0.9862	0.9391	-
	0.9358	0.90581	0.9743	0.9681	0.9860	0.9332	23.81%
UNet+ +	0.9078	0.8312	0.9503	0.8762	0.9418	0.8738	-
	0.9346	0.8773	0.9650	0.9089	0.9619	0.9108	47.06%
SegFormer	0.9127	0.8083	0.9431	0.8734	0.9159	0.8546	-
	0.9456	0.8681	0.9632	0.9302	0.9562	0.9045	∞
SegFormer (pretrained)	0.9583	0.9200	0.9783	0.9583	0.9584	0.9436	-
	0.9581	0.9195	0.9783	0.9622	0.9539	0.9434	25%
SegMenter	0.5289	0.3595	0.7386	0.4979	0.5641	0.3409	-
	0.8812	0.7876	0.9377	0.8747	0.8878	0.8390	∞
SegMenter (pretrained)	0.9473	0.8999	0.9725	0.9431	0.9516	0.9287	-
	0.9485	0.9021	0.9730	0.9418	0.9554	0.9303	0%

Bolded characters indicate accuracy data where the method in this paper has significantly improved compared to the original neural network

method, the optimal mDice of the final result is increased by 2.68%, and the optimal mIoU is increased by 4.61%. Compared with the optimal Recall, the optimal Precision is improved by more than 2%, and the optimal precision is improved by more than 3.27%. This phenomenon indicates that FP and FN metrics predicted by neural networks can be reduced using the proposed method, which conforms to the intuition exhibited during network training shown in Table 3, that is, an inside-out, universal to exact search strategy. Under the guidance of the SAA-SDM method, the convergence speed of UNet+ + network accuracy is increased by 47.06%. The proposed method is introduced into the UNet network structure. Although the optimal mDice of the final result is decreased by 0.3% and the optimal Recall is increased by 0.3%, the convergence speed of the network accuracy is improved by 42.86%.

In the framework of the most popular transformer type, our method can overcome the problem of slow improvement of training accuracy for large models to some extent. Specifically, in our experiments, under the SegFormer framework without pre-training parameters, the optimal mDice is increased by 3.75%, and the optimal mIoU is increased by 6% after adding the proposed method, which is close to the training results of the SegFormer framework loaded with pre-training parameters. In the framework of SegMenter without pre-trained parameters, the training speed of the network has been significantly improved after adding the method proposed in this paper. However, the accuracy can be further improved. Compared with the training results of the original SegMenter network framework, the optimal mDice is increased by 35.23%, and the optimal mIoU is increased by 42.81%. This result further demonstrates that the Transformer-type framework heavily depends on pre-trained parameters. In terms of accuracy, although there is no

further noticeable difference between the optimal mDice and the optimal mIoU of the SegFormer framework and the SegMenter framework loaded with pre-training data after adding the method described in this paper, their accuracy gaps are all controlled within 0.1%. However, regarding training accuracy convergence speed, the SegFormer framework loaded with pre-trained parameters shows a 25% speed improvement. Thus, the proposed method helps guide Transformer-type frameworks not loaded with pre-trained parameters to train in a favorable direction to achieve higher test accuracy.

Table 4 Test result data on pulmonary chest X-ray datasets. Each neural network has two rows of conclusion data. The first row of data corresponds to the original neural network framework proposed by the authors. In contrast, the second row of data represents the test results after incorporating our method SAA-SDM

Name	mIoU	Accuracy	Acceleration
Unet	0.9363	0.9699	-
	0.9334	0.9685	57.14%
SegNet	0.9175	0.9606	-
	0.9324	0.9679	86.67%
UNet+ +	0.8829	0.9693	-
	0.8934	0.9683	33.33%
UNet3 +	0.8989	0.9577	-
	0.9092	0.9629	0%
U2Net	0.8654	0.9339	-
	0.8793	0.9412	11.11%
SegFormer	0.8415	0.9387	-
	0.8883	0.9583	∞
SegMenter	0.8251	0.9130	-
	0.8372	0.9192	29.41%

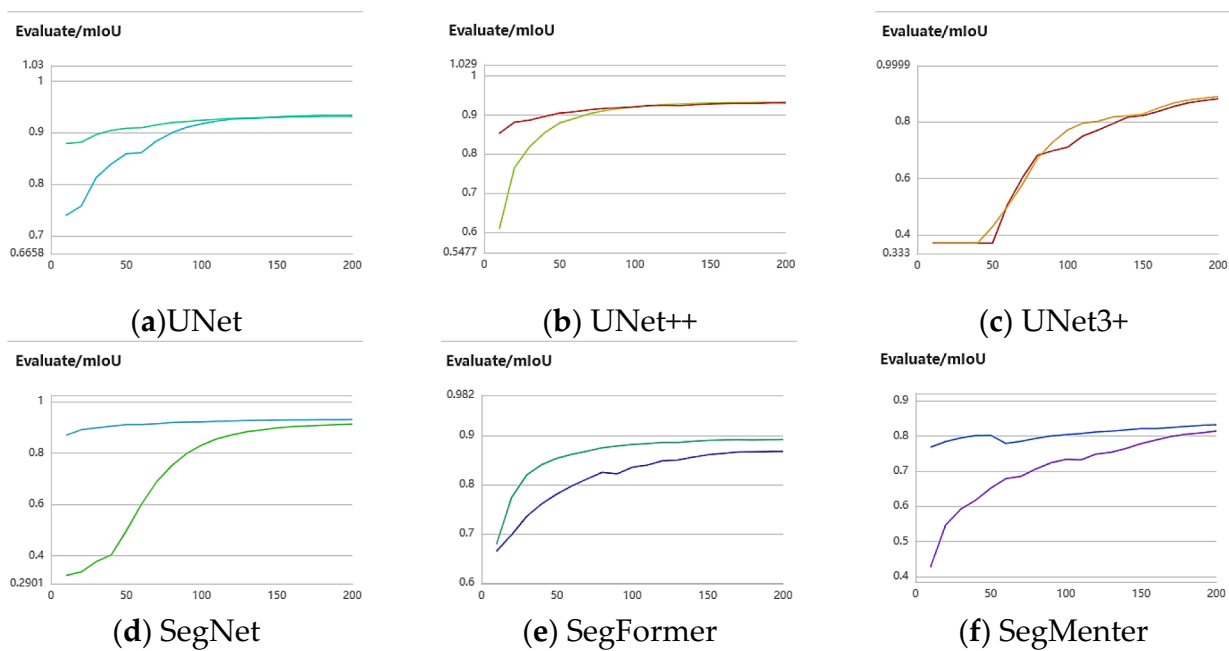


Fig. 3 Here are the test curves of different networks on the pulmonary chest X-ray datasets dataset. In all the line plots depicting images, the horizontal axis represents the training iteration, while the vertical axis represents the mean Intersection over Union (mIoU). In **a**, the green is SAA+UNet, and the blue is UNet; **b**

the red is SAA+UNet++, the green is UNet++; **c** the yellow is SAA+UNet3+ and the red is UNet3+; **d** the blue is SAA+SegNet, the green is SegNet; **e** the green is SAA+SegFormer, the purple is SegFormer; **f** the blue is SAA+SegMenter, and the purple is SegMenter

Table 4 shows our specific results on the pulmonary chest X-ray dataset. After adding the SAA-SDM method to the network, the accuracy is improved by 1.49% based on the SegNet neural network framework. Under the training of the other CNN-based U-network frameworks, the SAA-SDM method achieves more than 1% improvement in the optimal mIoU. It has a significant improvement in accuracy convergence speed. Similar to the training results on the TRUS dataset, in the Transformer-type framework, the boost after adding the SAA-SDM method is very significant, and we will not repeat it here.

We provide detailed curves of the test results for the interrupt test when trained in the pulmonary chest X-ray dataset, as shown in Fig. 3. Subgraphs (a),(b),(d),(f) all convey the same information to us, that is, with the addition of the proposed module, the network has perfect confidence to determine the target from the beginning. After simple and fast learning, NN learning focuses on refining the object boundary, as shown in the last two examples of Fig. 3. Moreover, Fig. 3e reflects the good guidance of the proposed module for the Transformer-type framework. The accuracy curve of the initial SegFormer network rises very slowly and converges to lower accuracies. However, a steeper accuracy curve is observed when the SegFormer framework is added to the training of the module in this paper. In contrast, a sharp rise process quickly reaches the convergence interval. Unfortunately, although the proposed method does not degrade the final accuracy of UNet3+ and drag down the training process, we do not observe a better response and boost in Fig. 3c.

Statistical Analysis

In order to obtain more robust conclusions, we conducted Brown-Forsythe tests and Mann-Whitney U tests on the acceleration performance of two convolutional neural networks, UNet and UNet++, and the mean Intersection over Union (mIoU) accuracy metric for the Segformer neural network without pre-trained parameters. These tests were conducted to compare the effectiveness of our method. First, we needed to verify whether adding SAA-SDM to the neural network's training process had a substantial impact. Subsequently, we assessed the magnitude of this impact.

The data used for the statistics are reported in Table 5. Based on Table 5, we subjected the data to Brown-Forsythe tests and Mann-Whitney U tests, and the conclusions are summarized in Table 6. For the two sets of UNet data without mean alignment, the Brown-Forsythe Test yielded a p -value of 1, which is greater than 0.05, leading to the acceptance of the null hypothesis that the two data groups have homogeneity of variance. The Mann-Whitney U test resulted in a p -value of 0.0112, less than 0.05, confirming that the two data groups do not come from the same distribution. This result indicates that the SAA-SDM module significantly affects the convergence speed of UNet during training.

After mean alignment, the Brown-Forsythe Test yielded a p -value of 0.2596, more significant than 0.05, allowing

Table 5 Six experiments were conducted on networks with SAA-SDM additions and their original networks, recording the number of iters required to reach the convergence state. All data in the table were scaled down by a factor of 10 to ensure the continuity of integers

Datasets	Methods	Num of test						Mean value	MV rate
		1	2	3	4	5	6		
TRUS	UNet (ours)	8	7	6	7	6	7	6.5	0.55
	UNet	13	12	14	12	12	12	12.5	
TRUS	UNet++(ours)	13	11	10	10	9	10	10.5	0.65
	UNet++	16	15	18	17	16	15	16	
Chest X-ray	UNet++(ours)	5	7	5	6	6	5	5.5	0.67
	UNet++	8	9	9	8	9	8	8.5	

Table 6 Data obtained from Table 5 underwent Brown-Forsythe tests and Mann–Whitney *U* tests, with corresponding statistics and *p*-values recorded

Group ID	Methods	Datasets	Data state	Brown-Forsythe test		Mann–Whitney <i>U</i> test	
				statis	<i>p</i>	statis	<i>p</i>
I.	UNet	TRUS	Before mean alignment	0	1	0	0.0040
			After mean alignment	1.4286	0.2596	13	0.3889
II.	UNet++	TRUS	Before mean alignment	0	1	0	0.0046
			After mean alignment	0.3226	0.5826	16.5	0.8580
III.	UNet++	Chest X-ray	Before mean alignment	1	0.3409	0	0.0041
			After mean alignment	1	0.3409	19.5	0.8586

us to accept the assumption of equal variances. Subsequently, the Mann–Whitney *U* test produced a *p*-value of 0.8325, more significant than 0.05, indicating that we cannot reject the null hypothesis that the two independent samples, UNet with speed enhancement and the original UNet, are from the same distribution.

Furthermore, we provide the boundary mean ratio values for *p*-values greater than 0.05, as shown in Table 7. This demonstrates the maximum acceptable error range and the boundary *U* statistics and *p*-values for rejecting the null hypothesis in the Mann–Whitney *U* test. Therefore, the conclusion can be drawn that for the three tested sample groups; their speed enhancement is at least 41.7%, 26.7%, and 22.2%, with corresponding average acceptable speed enhancements of 45.3%, 35.1%, and 33.3%. The maximum acceptable speed enhancements are 50%, 43.8%, and 44.4%.

To visually showcase the results generated by our approach, we present the segmentation results of UNet++ on the chest X-ray dataset. We overlay the segmentation

heatmap output onto the original images, as shown in Table 8.

Based on Table 8, the original UNet++ model shows many false positives (FP) in the prediction results obtained during the 50th round of training. These FP regions, indicated by red areas outside the green outlines, suggest that the network still needs to fully form the recognition of complete semantic information about the target object. This situation improves after the network undergoes training for 90 rounds, with a noticeable reduction in FP compared to previous data. In contrast, when testing with the neural network incorporating the SAA-SDM module, after just 50 rounds of training, there are hardly any distant FP errors in the predicted images. Similar conclusions can be drawn for other network architectures and test results, as explained and analyzed in “Ablation Experiments and Results”.

Next, we verify the accuracy improvement under the Segformer neural network framework. This experiment tests the image segmentation results obtained with the best test accuracy parameters after the same 200 training

Table 7 By adjusting the scaling factor, we searched for the minimum and maximum scaling factors that the Mann–Whitney *U* test could accept. The Brown-Forsythe test accepted all data in the table

Group ID	Mean value rate	Maximum acceptable rate	U-statistic	<i>p</i> value	Minimum acceptable rate	U-statistic	<i>p</i> value
I.	0.547	0.500	21	0.681	0.417	17	0.934
II.	0.649	0.438	29	0.089	0.267	7	0.089
III.	0.67	0.444	27	0.164	0.222	6	0.059

Table 8 This table displays the segmentation results of UNet++ on the chest X-ray dataset. “_ours” represents the output results of the neural network with the SAA-SDM module added. At the same time, “_num” indicates the segmentation results of the neural network at

the num-th round on the test dataset. In all images, the red portions represent the segmentation results provided by the neural network, while the green outlines correspond to the Ground Truth.

Dataset	Methods & iters	(a)	(b)	(c)	(d)	(e)
Chest X-Ray	UNetPP_50					
	_ours_50					
	UNetPP_90					

iters. Precision data is generated and statistically analyzed for mIoU accuracy following the same testing methodology as previous experiments.

Table 9 provides the mIoU test results for the loaded Segformer neural network framework with the SAA-SDM module. Accuracy tests were conducted on both the TRUS dataset and the chest X-ray dataset. Table 10 presents the statistics and *p*-values obtained from the Brown-Forsythe test and Mann–Whitney *U* test using the data from Table 9. All test results were not rejected in the Brown-Forsythe test, indicating

that they do not reject the assumption of homoscedasticity (equal variances). When comparing the mIoU metrics between the original Segformer framework and the Segformer framework with the added SAA-SDM module, the *U* test *p*-values were 0.0022 for both TRUS and chest X-ray datasets, which are smaller than 0.05. These results confirm a significant difference in mIoU accuracy after adding the SAA-SDM module to the Segformer framework following 200 rounds of training. Additionally, we performed mean alignment and searched for the rejection range, which revealed that after adding the

Table 9 Six experiments were conducted on the network with the addition of the SAA-SDM module and its original counterpart. After training completion, the experiments recorded the best mIoU on the

test set and calculated the mean mIoU of the six experiments, along with the ratio between the two means

Datasets	Methods	Num of test						Mean value	MV rate
		1	2	3	4	5	6		
TRUS	Segformer	0.81024	0.79920	0.80826	0.81552	0.81103	0.80494	0.8082	0.934
	_our	0.86922	0.86627	0.86805	0.86245	0.86168	0.86344	0.8652	
Chest X-ray	Segformer	0.85092	0.84949	0.85345	0.85585	0.84149	0.84889	0.8500	0.963
	_our	0.87493	0.88834	0.88948	0.88261	0.87617	0.88574	0.8829	

Table 10 Statistical analysis was performed on the data obtained in Table 9 using the Brown-Forsythe test and the Mann–Whitney *U* test. The table records the statistics and their corresponding *p*-values

Group ID	Datasets	Data state	Brown-Forsythe test		Mann–Whitney <i>U</i> test	
			statis	<i>p</i>	statis	<i>p</i>
IV.	TRUS	Before mean alignment	0.8263	0.3847	0	0.0022
		After mean alignment	1.0575	0.3280	19	0.9372
V.	Chest X-ray	Before mean alignment	0.7266	0.4140	0	0.0022
		After mean alignment	0.5863	0.4615	16	0.8182

Table 11 By adjusting the scaling factor, we searched for the minimum and maximum scaling factors that the Mann–Whitney *U* test could accept. The data recorded in this table were all accepted by the Brown-Forsythe test

Group ID	Mean value rate	Maximum acceptable rate	U-statistic	<i>p</i> value	Minimum acceptable rate	U-statistic	<i>p</i> value
IV.	0.934	0.0733	7	0.0931	0.0597	30	0.0649
V.	0.963	0.0459	6	0.0649	0.0289	30	0.0649

SAA-SDM module, the Segformer framework achieved a minimum accuracy improvement of at least 5.97% and 2.89% on the TRUS dataset, with an average improvement accepted by the *U* test of 6.6% and 3.7%, and a maximum improvement accepted by the *U* test of 7.33% and 4.59% (Table 11). We also provide some batch_cost and train_cost indicators during the

training process, showing almost complete overlap, indicating that our method added minimal burden to the network.

Figure 4 presents a detailed overview of resource consumption during the training process. It is essential to note that the primary purpose of this test was to assess the overall impact of loading SAA-SDM on the network’s

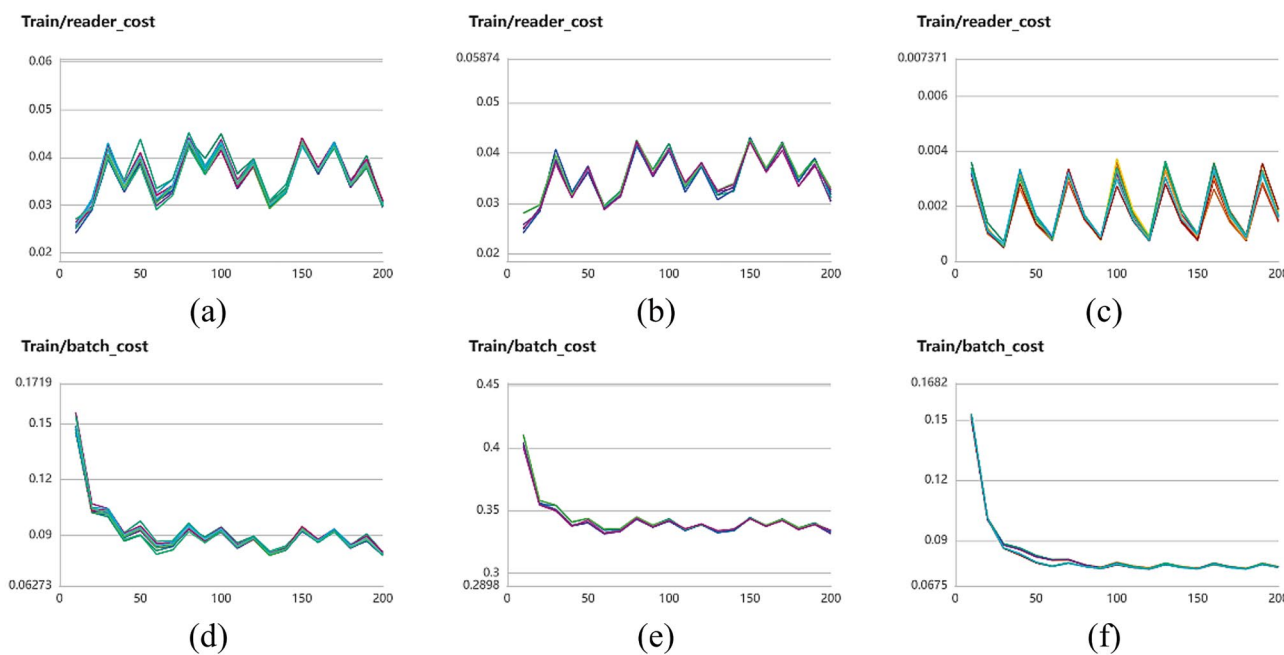


Fig. 4 The figure displays the time consumption for reading data and training a single network through multiple training sessions. The vertical axis represents time in minutes, and the horizontal axis represents

the iteration rounds. **a–c** depict line graphs showing the time consumption for data reading, while **d–f** show line graphs for training time per batch

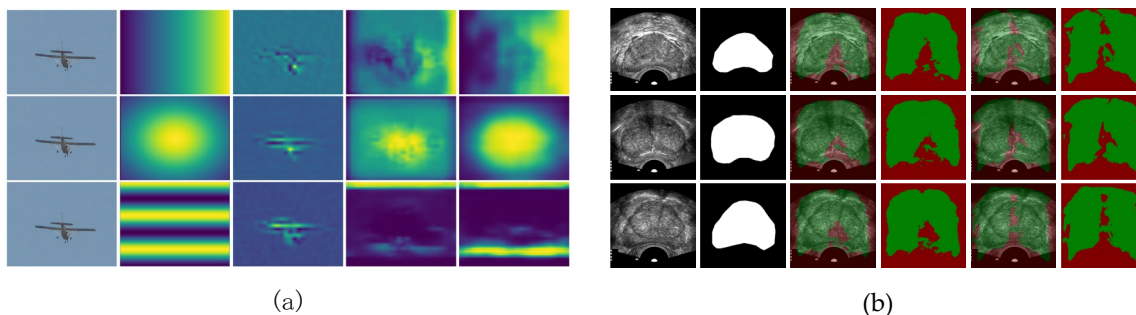


Fig. 5 **a** Represents experiments conducted by Islam et al. [13]. **b** With the data used in this paper. In **b**, the first column represents the input, the second column shows the Ground Truth, and columns three to six display the neural network’s prediction results

workload, independent of image size. Therefore, we adjusted the image scale to 64×64 in the experiments for quick testing. Any subgraph within Fig. 4 encompasses multiple sets of monitoring data for both loaded and unloaded SAA-SDM modules, and these data points are nearly superimposed in the schematic, indicating that our method added minimal burden to the network.

Conclusions

This paper proposes a data processing method that can accelerate the accuracy convergence of neural networks. Reasonable use of this method can make neural networks reach the accuracy convergence interval after less training and significantly shorten the training time. By verifying two 2D datasets, the proposed method has apparent effects under the U-shaped neural network based on CNN architecture and the neural network based on transformer architecture. In addition, experiments show that neural networks based on transformer architectures have much room for improvement without pre-trained parameters. The method proposed in this paper can guide this network training well and improve accuracy. In addition, in our study, the proposed method at least did not find the situation of impairing training; that is, it did not slow down the training process of the network, increasing the batch cost of the network and the weakening difference of the optimal mIoU accuracy of more than 0.5% compared with the initial network.

Moreover, our method can be easily extended to studying 3D images, as reported in the V-Net and 3D-Unet papers. Our research applies to solids with strong rigidity; the most typical example is the 2D section of the human organ and the 3D overall shape of the organ. At the same time, the proposed method is easy to combine with any current mainstream neural network and does not cause additional burdens on the network.

Discussion

Much of the research currently focuses on refining the effective parameter count within neural networks to accelerate their operation speed and reduce the time required for neural network predictions. Standard techniques include low-rank decomposition of network parameters, structural pruning of neural network architectures, and knowledge distillation. Reducing the space occupied by neural networks facilitates their efficient deployment on less powerful devices, making it easier to popularize neural network-assisted medical diagnostics. However, these methods primarily concentrate on optimizing parameters of neural networks that have been trained thousands of times. The training process of

neural networks often relies on high-performance computing devices. Since resolving the gradient vanishing problem by architectures like ResNet, neural network structures and hierarchies have become increasingly complex, leading to longer training times and greater computational demands. Unlike datasets for vehicle recognition and license plate detection tasks, medical image datasets are more challenging to collect and annotate. This case poses obstacles to the practical training of neural networks in the medical field.

In the study by [13], a phenomenon was introduced: even when there is no direct correlation between the Ground Truth and the image, neural networks are still capable of extracting elements from the image to a certain extent, as illustrated in Fig. 5. This suggests that the original information in the image has the potential to be retained and is not entirely influenced by the Ground Truth. Biological organs often exhibit relatively fixed sizes, shapes, and positions, which are prerequisites for traditional organ segmentation methods to achieve high-precision segmentation. Using unaltered medical organ images to train neural networks is akin to relinquishing some advantages of the target, which is unwise. We conducted similar experiments by extending the dataset used in this study. We made predictions with a neural network trained on the TRUS dataset but loaded with an SAA-SDM module created using the chest X-ray dataset, as shown in Fig. 5b. This finding is consistent with the conclusions drawn in [13] in the abovementioned experiments. Therefore, providing semantic information to the network is feasible for faster learning. The proposal offers a rapid model training approach for researchers in medical diagnostics and treatment, helping them avoid inefficient and lengthy training processes that consume valuable resources and time.

However, this research still has some limitations. The SAA-SDM module assumes that the target objects have similar sizes, shapes, and positions. Although this study improved the module's generalization performance through data transformation and augmentation after removing the SAA-SDM module, the proposed method may not be suitable for tasks involving targets with significant shape variations, such as thyroid nodule segmentation in ultrasound images and glioma segmentation in brain MRI images. Therefore, further research is needed to enhance the module's robustness under these three conditions, such as endowing the SAA-SDM module with deformability through learnable affine matrices to broaden its applicability and strengthen its versatility.

Author Contribution Conceptualization, Y.S. and C.G.; methodology, C.G.; software, S.Y. and C.G.; validation, C.G., R.Z., S.Y. and B.L.; formal analysis, C.G.; investigation, S.Y. and C.G.; data curation, Y.S.; writing—original draft preparation, S.Y. and C.G.; writing—review and editing, C.G.; visualization, S.Y. and C.G.; supervision, Y.S. and B.L.; project administration, Y.S. and B.L. All authors have read and agreed to the published version of the manuscript.

Data Availability Published public datasets were used for our experiments.

Declarations

Informed Consent Not applicable.

Conflict of Interest The authors declare no competing interests.

References

- Hao, S.; Zhou, Y.; Guo, Y. A Brief Survey on Semantic Segmentation with Deep Learning. *Neurocomputing* 2020, 406, 302–321.
- Asgari Taghanaki, S.; Abhishek, K.; Cohen, J.P.; Cohen-Adad, J.; Hamarneh, G. Deep Semantic Segmentation of Natural and Medical Images: A Review. *Artificial Intelligence Review* 2021, 54, 137–178.
- Hodge, A.C.; Fenster, A.; Downey, D.B.; Ladak, H.M. Prostate Boundary Segmentation from Ultrasound Images Using 2D Active Shape Models: Optimisation and Extension to 3D. *Computer Methods and Programs in Biomedicine* 2006, 84, 99–113, <https://doi.org/10.1016/j.cmpb.2006.07.001>.
- Wang, X.-F.; Min, H.; Zou, L.; Zhang, Y.-G.; Tang, Y.-Y.; Chen, C.-L.P. An Efficient Level Set Method Based on Multi-Scale Image Segmentation and Hermite Differential Operator. *Neurocomputing* 2016, 188, 90–101.
- Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE transactions on neural networks and learning systems* 2021.
- Gunel, B. *Leveraging Prior Knowledge and Structure for Data-Efficient Machine Learning*; Stanford University, 2022;
- Li, J.; Nebelung, S.; Schock, J.; Rath, B.; Tingart, M.; Liu, Y.; Sirros, N.; Eschweiler, J. A Novel Combined Level Set Model for Carpus Segmentation from Magnetic Resonance Images with Prior Knowledge Aligned in Polar Coordinate System. *Computer Methods and Programs in Biomedicine* 2021, 208, 106245.
- Peng, T.; Wu, Y.; Qin, J.; Wu, Q.J.; Cai, J. H-ProSeg: Hybrid Ultrasound Prostate Segmentation Based on Explainability-Guided Mathematical Model. *Computer Methods and Programs in Biomedicine* 2022, 219, 106752.
- Zhang, Z.; Gao, S.; Huang, Z. An Automatic Glioma Segmentation System Using a Multilevel Attention Pyramid Scene Parsing Network. *Current Medical Imaging* 2021, 17, 751–761.
- Long, X.; Zhang, W.; Zhao, B. PSPNet-SLAM: A Semantic SLAM Detect Dynamic Object by Pyramid Scene Parsing Network. *IEEE Access* 2020, 8, 214685–214695.
- Cheng, Y.; Wang, D.; Zhou, P.; Zhang, T. Model Compression and Acceleration for Deep Neural Networks: The Principles, Progress, and Challenges. *IEEE Signal Processing Magazine* 2018, 35, 126–136, <https://doi.org/10.1109/MSP.2017.2765695>.
- Lebedev, V.; Lempitsky, V. Speeding-up Convolutional Neural Networks: A Survey. *Bulletin of the Polish Academy of Sciences: Technical Sciences* 2018, 66, 799–810, <https://doi.org/10.24425/bpas.2018.125927>.
- Islam*, M.A.; Jia*, S.; Bruce, N.D.B. How Much Position Information Do Convolutional Neural Networks Encode?; September 25 2019.
- Liu, R.; Lehman, J.; Molino, P.; Petroski Such, F.; Frank, E.; Sergeev, A.; Yosinski, J. An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution. *Advances in Neural Information Processing Systems* 2018, 31.
- Li, S.; Zhang, C.; He, X. Shape-Aware Semi-Supervised 3D Semantic Segmentation for Medical Images. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*; Springer, 2020; pp. 552–561.
- Liu, S.; Li, Y.; Li, X.; Cao, G. Shape-Aware Multi-Task Learning for Semi-Supervised 3D Medical Image Segmentation. In *Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; December 2021; pp. 1418–1423.
- Grandvalet, Y.; Bengio, Y.; Chapelle, O.; Schölkopf, B.; Zien, A. *Entropy Regularization*. Springer 2006.
- Lee, D.-H. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *Proceedings of the Workshop on challenges in representation learning, ICML; 2013; Vol. 3*, p. 896.
- Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *IEEE 2016*, <https://doi.org/10.1109/3DV.2016.79>.
- Wu, P.; Liu, Y.; Li, Y.; Shi, Y. TRUS Image Segmentation with Non-Parametric Kernel Density Estimation Shape Prior. *Biomedical Signal Processing & Control* 2013, 8, 764–771, <https://doi.org/10.1016/j.bspc.2013.07.002>.
- Jaeger, S.; Karargyris, A.; Candemir, S.; Folio, L.; Siegelman, J.; Callaghan, F.; Xue, Z.; Palaniappan, K.; Singh, R.K.; Antani, S. Automatic Tuberculosis Screening Using Chest Radiographs. *IEEE Transactions on Medical Imaging* 2014, 33, 233–245, <https://doi.org/10.1109/TMI.2013.2284099>.
- Hooda, R.; Mittal, A.; Sofat, S. Lung Segmentation in Chest Radiographs Using Fully Convolutional Networks. *Turkish Journal of Electrical Engineering and Computer Sciences* 2019, 710–722, <https://doi.org/10.3906/elk-1710-157>.
- Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*; Springer, 2015; pp. 234–241.
- Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. 2018, https://doi.org/10.1007/978-3-030-00889-5_1.
- Huang, H.; Lin, L.; Tong, R.; Hu, H.; Wu, J. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. *arXiv* 2020, <https://doi.org/10.1109/ICASSP40776.2020.9053405>.
- Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Jagersand, M. U2-Net: Going Deeper with Nested U-Structure for Salient Object Detection. *Pattern Recognition* 2020, 106, 107404, <https://doi.org/10.1016/j.patcog.2020.107404>.
- Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. 2017; <https://doi.org/10.17863/CAM.17966>.
- Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B. Attention U-Net: Learning Where to Look for the Pancreas. 2018, <https://doi.org/10.48550/arXiv.1804.03999>.
- Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. 2021, <https://doi.org/10.48550/arXiv.2105.15203>.
- Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation.; 2021; pp. 7262–7272.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.