Biomedical Signal Processing and Control 94 (2024) 106254

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Improving deep learning based segmentation of scars using multi-view images

Jian Zhou [a], Yuqing Dai [a], Dongmei Liu [b], Weifang Zhu [a], Dehui Xiang [a], Xinjian Chen [a,d], Fei Shi [a,c,*], Wentao Xia [b,**]

[a] The MIPAV Lab, The School of Electronics and Information Engineering, Soochow University, 1 Shizi Street, Suzhou, 215006, China
[b] Shanghai Key Laboratory of Forensic Medicine, Shanghai Forensic Service Platform, Academy of Forensic Science, Ministry of Justice, 1347 West Guangfu Road, Shanghai, 200063, China
[c] The Fariver Innovation Technologies Company, Ltd., 2000 Majian Road, Suzhou, 215151, China
[d] The State Key Laboratory of Radiation Medicine and Protection, Soochow University, 199 Renai Road, Suzhou, 215123, China

ARTICLE INFO

Keywords:
Skin scar
Deep learning
Image segmentation
Multi-view co-segmentation
Data augmentation

ABSTRACT

The utilization of deep learning for scar segmentation in photographs enables automated and non-contact quantitative analysis of skin scars. Meanwhile, multi-view photographs are commonly employed to capture the 3D information of scars. In this paper, we propose a two-stage deep learning based segmentation framework for delineating scars from surrounding skin, leveraging multi-view images to achieve enhanced segmentation results compared to single-view approaches. In the first stage, a data augmentation method based on 3D reconstruction and view interpolation is proposed. The generated images are used in a semi-supervised setting to train a single-view segmentation network. In the second stage, a multi-view co-segmentation network (MVCSNet) is proposed to exploit the mutual information between views and to further refine the segmentation. The multi-view feature interaction module (MVFI) uses the prior segmentation results from the first stage, computes feature similarities across views, and optimizes the features. The proposed method was evaluated on two multi-view image datasets containing linear scars and patchy scars, respectively. The results show that the proposed data augmentation method can improve the generalization of the model, particularly for the dataset with smaller size. Comparative analyses demonstrate the superior performance of MVCSNet over other deep learning based segmentation or co-segmentation algorithms.

1. Introduction

Scars are the result of the skin's natural healing process following injury, characterized by an excessive accumulation of collagen fibers and alterations in the tissue structure and pigmentation [1]. The quantitative assessment of scars is essential in both forensic investigations and clinical dermatology. In forensics, scar measurements can be used to assess the severity of human injuries and to infer the injury time and vulnerant, so as to assist crime investigation and to protect the legitimate rights and interests of victims [2,3]. In dermatology, scar measurements can be utilized to monitor the skin healing process and to evaluate the treatment outcomes [4–6]. Utilizing image-based quantitative analysis for scar assessment offers a non-invasive approach that allows for accurate and easily recordable measurement results. Additionally, the implementation of automatic image segmentation enhances the efficiency and objectivity of the analysis process. However, due to the varied forms and sizes of scars, automatic analysis of scars remains a challenging problem. Therefore, this paper aims to develop an automatic framework for scar image segmentation with improved accuracy, which facilitates the subsequent quantitative analysis.

Some methods were proposed for the similar task of skin wound analysis, using either traditional segmentation methods [7–9] or deep learning frameworks [10–12]. These studies primarily focused on segmenting wounds in 2D images captured from a frontal view. However, since scars and wounds are typically adhered to the curved surfaces of the human body, a single-view image may not provide sufficient spatial information. In the absence of specialized equipment, multi-view photogrammetry [3] can be employed, which involves taking images of the scar from multiple viewpoints and then reconstructing a 3D model. This not only allows for measurement in 3D space, but also provides more comprehensive information for segmentation than in the

* Corresponding author at: The MIPAV Lab, The School of Electronics and Information Engineering, Soochow University, 1 Shizi Street, Suzhou, 215006, China.
** Corresponding author at: Shanghai Key Laboratory of Forensic Medicine, Shanghai Forensic Service Platform, Academy of Forensic Science, Ministry of Justice, 1347 West Guangfu Road, Shanghai, 200063, China.
E-mail addresses: shifei@suda.edu.cn (F. Shi), xiawt@ssfjd.cn (W. Xia).

https://doi.org/10.1016/j.bspc.2024.106254
Received 22 September 2023; Received in revised form 23 February 2024; Accepted 20 March 2024
Available online 28 March 2024

single-view setting. Some prior research endeavors have incorporated multi-view images into the segmentation process. Wannous et al. [13] conducted skin area segmentation and wound tissue type classification on individual images, subsequently merging the classification results based on the reconstructed 3D model. Liu et al. [14] applied the least squares conformal mapping algorithm to unfold the 3D model into a 2D image, and then segmented the wound with an interactive method. Niri et al. [15] proposed a data augmentation method for deep learning based on the reconstructed 3D wound model. From multi-view images, the optimal view was selected, and its segmentation result was projected to other views as ground truth. These studies showed that information from multi-view images could improve the segmentation performance, but they each only investigated the multi-view information from a single perspective. In this paper, we propose a framework that can take more advantage of the abundant information available in multi-view images, where both the 3D spatial information and the inter-image feature resemblance are exploited.

In recent years, medical image segmentation methods based on convolutional neural networks (CNNs) have been widely studied and applied [16,17]. UNet [18] is the classic CNN in the field of medical image segmentation, and many other network structures were proposed on this basis, with enhanced ability of feature extraction [19,20]. CNNs can achieve high performance with short inference time, which greatly facilitates practical applications. However, their performance depends on the quantity and quality of training data. In the case of insufficient data, it is difficult to obtain a robust model [21]. To address the issue of poor model generalization caused by insufficient data, data augmentation techniques are commonly employed to expand the training dataset. Conventional data augmentation techniques typically involve random transformations such as rotation, flipping, and contrast enhancement [22–24]. Additionally, for medical images where manual labels are unavailable for some data, the semi-supervised learning strategy can be applied. Usually, a teacher model is first trained with labeled data, which is then applied on the unlabeled data to generate pseudo-labels, and finally all the data is mixed to retrain a final model [25–27]. Therefore, to improve the generalization ability of the segmentation model, we propose a data augmentation method based on the reconstructed 3D model, and the semi-supervised learning strategy is adopted for the augmented data.

Object co-segmentation refers to segmenting objects of the common category from a set of images. In deep learning based co-segmentation methods, the deep features are optimized by jointly utilizing the information from multiple images. For example, SAAB [28] used the channel-wise attention module, placed in the bottleneck layer of the network, to select semantically related features from a pair of images. DOCS [29] used a mutual correlation layer to perform feature matching to obtain correspondence maps that were fused with other features to help co-segmentation. COSNet [30] proposed a co-attention module to compute the attention summaries that encoded the correlations between features. In this paper, we design a multi-view co-segmentation network (MVCSNet) to achieve consistency optimization based on feature interactions.

In summary, we are committed to improving CNN-based scar segmentation results by effectively utilizing multi-view image data. We propose a two-stage framework for scar segmentation. In each stage, the multi-view images are utilized in different ways. The contributions of the paper are listed as follows:

- In the first stage, we propose a data augmentation method based on view interpolation. After 3D reconstruction from multi-view images, the camera parameters of different views are interpolated. With the resulted new parameters, the 3D model is projected onto the 2D plane to generate more camera views. Then the new views are used in a semi-supervised learning setting to further optimize the segmentation network.

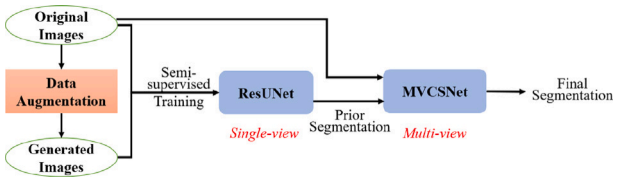

Fig. 1. The flowchart of the proposed scar segmentation framework.

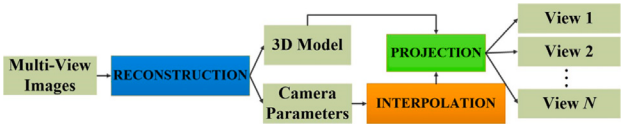

Fig. 2. Data augmentation process based on 3D view interpolation.

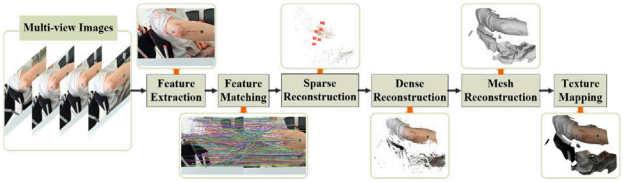

Fig. 3. 3D reconstruction process based on SfM algorithm.

- In the second stage, we propose a multi-view co-segmentation network. A multi-view feature interaction (MVFI) module is embedded in the middle of the network for feature optimization across views. Guided by the prior segmentation results, the MVFI module captures the category features of each view, and then performs consistency optimization on features of other views.
- The proposed method is evaluated on two datasets with two types of scars respectively and achieved superior performance than some single-view based segmentation methods and some co-segmentation methods.

2. Methods

Fig. 1 shows the flowchart of the proposed scar segmentation framework. This section first introduces the proposed data augmentation method based on 3D view interpolation and the semi-supervised learning strategy for the first-stage single-view segmentation, then introduces the overall structure of the MVCSNet and the proposed MVFI module for the second-stage multi-view co-segmentation, and finally gives the loss function used.

2.1. Data augmentation based on 3D view interpolation

The proposed data augmentation process is an offline data augmentation method. As shown in Fig. 2, first, we use the SfM [31] algorithm to perform 3D reconstruction on multi-view images to obtain 3D models and camera parameters for each view. Then, we interpolate the sequence of camera parameters. Finally, the interpolated camera parameters are applied to the 3D model to generate new 2D camera views. The proposed data augmentation method not only increases the amount of training data, but also improves the ability of the model to perceive objects from different views.

The SfM [31] is a prevalent algorithm for 3D reconstruction based on multi-view images. As shown in Fig. 3, the overall process of SfM mainly includes feature extraction, feature matching, sparse reconstruction, dense reconstruction, mesh reconstruction, and texture mapping.

Besides the 3D model, the camera parameters corresponding to each view, including intrinsic and extrinsic ones, can be estimated by

Table 1
Size of the dataset.

Numbers	Linear scar dataset					Patchy scar dataset		
	fold1	fold2	fold3	fold4	fold5	fold1	fold2	fold3
Original images	149	148	149	149	149	52	54	53
Scar samples	24	26	27	24	23	6	5	7
Generated images	226	256	234	190	226	47	67	58

the SfM algorithm. The extrinsic camera parameters are the rotation and translation of each viewpoint. We propose to interpolate these parameters to generate new viewpoints. The 3D rotation parameters of the ith view can be represented by a quaternion, which is defined as

$$\vec{q}_i = [\cos\frac{\theta_i}{2}, u_{ix}\sin\frac{\theta_i}{2}, u_{iy}\sin\frac{\theta_i}{2}, u_{iz}\sin\frac{\theta_i}{2}], i = 1, 2, \ldots, N \tag{1}$$

where $\vec{u}_i = [u_{ix}, u_{iy}, u_{iz}]$ represents the unit-length rotation axis, and θ_i represents the rotation angle. N represents the total number of views. Quaternions can be interconverted with rotation matrices.

Assuming there are two views with quaternions $\vec{q}_n$ and $\vec{q}_{n+1}$, and translation vectors $\boldsymbol{t}_n$ and $\boldsymbol{t}_{n+1}$, linear interpolation can be used to obtain the parameter of a new viewpoint as follows.

$$\vec{q}_{n\lambda} = (1-\lambda)\vec{q}_n + \lambda\vec{q}_{n+1} \tag{2}$$

$$\boldsymbol{t}_{n\lambda} = (1-\lambda)\boldsymbol{t}_n + \lambda\boldsymbol{t}_{n+1} \tag{3}$$

where $\lambda \in (0, 1)$ is the interpolation ratio. For each new viewpoint, its rotation matrix $\boldsymbol{R}_{n\lambda}$ is calculated from the generated $\vec{q}_{n\lambda}$.

$$\boldsymbol{R}_{n\lambda} = \begin{pmatrix} 1-2q_2^2-2q_3^2 & 2q_1q_2+2q_0q_3 & 2q_1q_3+2q_0q_2 \\ 2q_1q_2-2q_0q_3 & 1-2q_1^2-2q_3^2 & 2q_2q_3+2q_0q_1 \\ 2q_1q_3+2q_0q_2 & 2q_2q_3-2q_0q_1 & 1-2q_1^2-2q_2^2 \end{pmatrix} \tag{4}$$

where q_0, q_1, q_2, q_3 are the four elements in $\vec{q}_{n\lambda}$. The new rotation matrix and translation vector are used to project the 3D model into a new generated image. Specifically, the texture of the 3D model is mapped to a 2D plane by mapping the 3D coordinates $[x, y, z]$ to the 2D coordinates $[x', y']$ by

$$[x', y'] = K[\boldsymbol{R}|\boldsymbol{t}][x, y, z] \tag{5}$$

where K represents the matrix of intrinsic camera parameters, and $[\boldsymbol{R}|\boldsymbol{t}]$ represents the matrix of extrinsic camera parameters.

In this paper, we insert a new viewpoint between two consecutive views using $\lambda = 0.5$. Besides, the parameters of each original view are also used to generate a projection. This results in a image that is slightly different than the original image, and can be seen as a disturbance of the data. Therefore, the intended number of generated images are twice of the original ones. However, for some samples with few views or poor image quality, the generated images may exhibit reconstruction errors, and in some cases generated views may only contain a small part of the scar. In such scenarios, the generated images are discarded. Therefore the amount of augmented images actually used is smaller. The number of generated images included in the augmented dataset is shown in Table 1.

2.2. Semi-supervised learning with augmented data

As ground truth labels are unavailable for the generated images, we adopt the self-training strategy [25], which is commonly used in semi-supervised learning, for the first-stage single-view segmentation network. As shown in Fig. 4, first the segmentation network is trained on the original data. Then the generated images are fed into the trained network, and pseudo-labels are obtained. Finally the generated images with pseudo-labels and the original data with ground truth labels are mixed to form the augmented dataset, with which the segmentation network is retrained.

A ResUNet [32] is used as the single-view segmentation network. The encoder is the ResNet34 [33] model pre-trained on ImageNet, and the decoder consists of convolution blocks and upsampling.

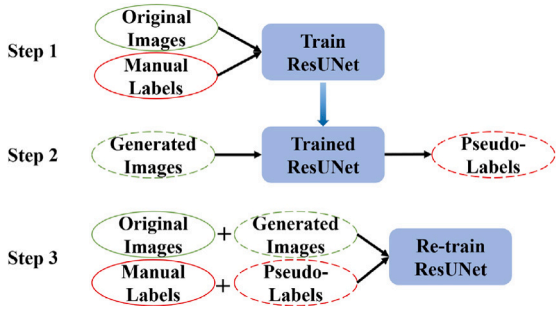

Fig. 4. Steps of semi-supervised learning with augmented data.

2.3. Multi-view co-segmentation network

After a prior segmentation is obtained by the ResUnet trained with the augmented data, a multi-view co-segmentation network (MVCSNet) is constructed to further refine the segmentation by considering the consistency among multi-view images of the same scar sample. As shown in Fig. 5, the MVCSNet is an encoder–decoder network with skip connections. The encoder is modified upon the ResNet34 [33]. Considering that for thin linear scars and low-contrast patchy scars, detail information will be lost on low resolution levels, the first maximum pooling layer is removed. The fourth residual block is also discarded to keep a high resolution while reducing model complexity. Therefore, the encoder of MVCSNet downsamples the input image three times. The decoder includes cascaded convolution and upsampling blocks. The MVFI module is inserted in the middle of the encoder and decoder. It uses the prior segmentation results to guide the optimization of multi-view features.

In the co-segmentation process, different views of each sample are input into the MVCSNet together, and their segmentation results are output together. The view dimension and batchsize dimension are merged before the encoder, to ensure that all views are encoded together in a batch. The features of each view are split after the encoder, as they need to be processed separately in the MVFI module. Similar merging and splitting is performed for the decoder, where all views are decoded together and then separated for output. Specifically, let the original input image be $I \in R^{B\times3\times N\times H\times W}$, where B is the batchsize, N is the number of views, H is the image height, and W is the width. $X \in R^{BN\times3\times H\times W}$ is the result of view merging on I. X is fed into the encoder to obtain a deep feature F. View separation is performed on F in to obtain $F_i \in R^{B\times c\times h\times w}, i = 1\cdots N$, where c, h, and w is channel number, height, and width of the features. The separated feature F_i is input into the MVFI module to get the optimized features $F_i' \in R^{B\times c\times h\times w}$. Subsequently, $F' \in R^{BN\times c\times h\times w}$ is obtained by view merging. The decoder takes F' as the input and outputs $O \in R^{BN\times1\times H\times W}$. Finally, view separation is performed on O to get segmentation results $O_i \in R^{B\times1\times H\times W}$ for each view in the batch. Note that N can be different for each scar sample, and therefore the input size of the encoder and decoder are designed to be variable.

2.4. Multi-view feature interaction module

This subsection introduces the multi-view feature interaction (MVFI) model used in the MVCSNet. In MVFI, F_i, the features from each

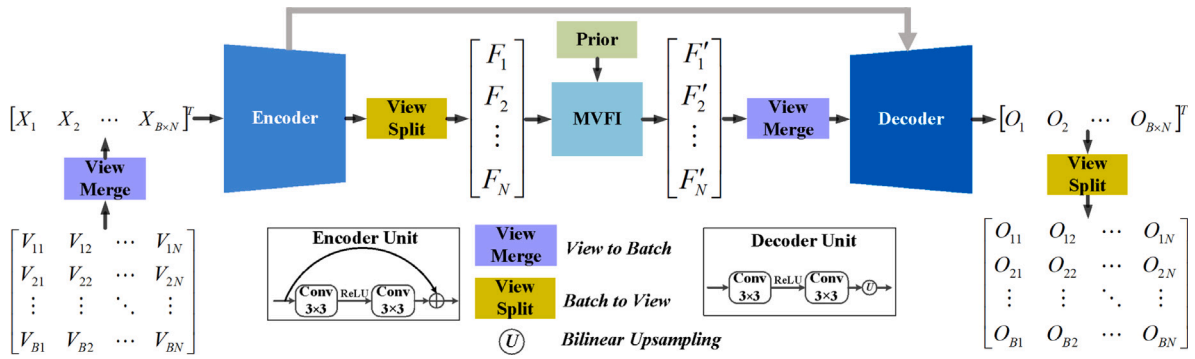

Fig. 5. Multi-view co-segmentation network.

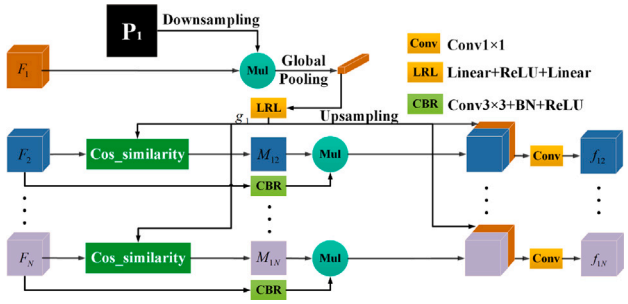

Fig. 6. Multi-view feature interaction module.

view, with its segmentation prior P_i, is used to optimize the features from all other views. Correspondingly, the features from each view are optimized multiple times, and these optimized features are fused to give the output feature for this view. Without loss of generality, Fig. 6 shows the structure where F_1 and P_1 are used as the reference to optimize all other views.

The cosine similarity defined in (6) is used in MVFI to measure the consistency between features from different views.

$$\text{cos\_similarity} = \frac{g_x \cdot g_y}{\|g_x\| \ \|g_y\|} \tag{6}$$

where g_x and g_y are two feature vectors.

As in Fig. 6, assuming the first view as the reference, the prior segmentation result P_1, is down-sampled to the size of F_1, which is $B \times c \times h \times w$, and then multiplied with F_1 to obtain a masked feature. Then global average pooling followed by cascaded fully connected layers is used to obtain $g_1 \in R^{B \times c \times 1 \times 1}$. Since the background is removed by masking, g_1 represents the category feature of the targeted scar area. Based on this, the similarity feature map M_{1j} is obtained by calculating the cosine similarity between g_1 and each pixel position of $F_j, j = 2 \cdots N$.

$$M_{1j}(p,q) = \text{cos\_similarity}(g_1, F_j(p,q)), j = 2, \ldots, N \tag{7}$$

F_j, after some regular trainable processing, is multiplied by its corresponding similarity matrix M_{1j}, so that the high similarity parts of the feature are boosted and the low similarity parts are suppressed. Finally, the category feature g_1, which contains important global semantic information, is used again to further optimize the features. Specifically, after upsampling, it is fused with the features through concatenation and 1×1 convolution. The final output f_{1j} is restored to the same size as the input F_j.

Such operation is repeated using other F_i and P_i as references, respectively. Finally, a set of features $f_{ij}, i, j = 1 \cdots N, i \neq j$ can be obtained, which represents the optimized output of the jth view, where the ith view is used as a reference. For each view, the features resulted from optimization by all other views are then summed to get the final output features of MVFI.

$$F'_j = \sum_{i=1, i \neq j}^{N} f_{ij} \tag{8}$$

In MVFI, each view uses the rest $N-1$ views for consistency optimization, and finally the $N-1$ optimization results are integrated. Such an ensemble operation can reduce the impact of errors in prior segmentation and improve the reliability of feature optimization.

2.5. Loss function

For both first-stage single-view segmentation and second-stage multi-view segmentation, we adopt the joint loss function of focal loss [34] and Dice loss.

$$L_{total} = L_{focal} + L_{Dice} \tag{9}$$

The focal loss [34] function is defined as

$$L_{focal} = -\sum_i (1-\hat{p})^{\gamma} y_i \log(\hat{p}_i) - \hat{p}_i^{\gamma}(1-y_i)\log(1-\hat{p}_i) \tag{10}$$

where $\hat{p}_i$ is the predicted value of the ith pixel, y_i is the ground truth value of the ith pixel, and γ is the adjustment factor, which is set to 2 in our experiments. The Dice loss function is defined as

$$L_{Dice} = 1 - \frac{2\sum_i \hat{p}_i y_i + \epsilon}{\sum_i \hat{p}_i^2 + \sum_i y_i^2 + \epsilon} \tag{11}$$

where ϵ is a small smoothing factor.

3. Experimental settings

3.1. Datasets

The data used in the experiments are clinical data collected at the Academy of Forensic Science, Ministry of Justice, Shanghai, China. The collection and analysis of image data were approved by the Institutional Review Board of the Academy of Forensic Science. A smartphone was used to collect 744 images from 130 linear scars samples, and 159 images from 18 patchy scars samples. Each sample had 3 to 10 views. All images were taken by the same device with a resolution of 3456 × 4608. The scar region in each image was manually annotated by a forensic expert using the LabelMe software [35]. Dense points were placed along the edges of scars and connected to form a mask for the scar regions. The linear scars and patchy scars differed greatly in shape, and therefore they were treated as two independent datasets, and training and testing are performed separately.

Due to the small amount of data, we use cross-validation in all experiments. Five-fold cross validation was performed on linear scars and three-fold cross validation on patchy scars. The amount of data per fold is shown in Table 1.

3.2. Implementation details

The segmentation experiments were performed on the public platform PyTorch with NIVIDIA GeForce RTX3090 graphics card and 24G video memory. The implementation of SfM [31] algorithm was based on Colmap [36] and OpenMVS [37] open source library.

The hyperparameter settings for network training were as follows. For ResUNet, the number of epochs was set to 100, and the batchsize was set to 4. Poly learning rate policy was adopted, and the initial learning rate is 0.01. The optimizer used was SGD with a momentum of 0.9 and weight decay of 0.0001. For MVCSNet, the number of epochs was set to 60, and the batchsize was set to 1 due to restraints of computational complexity. Other settings were the same.

In 3D reconstruction by SfM, the input images were downsampled to 864×1152. The input to the networks was rescaled to 512×512. In order to improve the robustness of the model, online data enhancement was employed during training, including random flipping and rotation. In the multi-view co-segmentation experiment, too many views would result in a huge amount of calculation. Therefore, the maximum number of views was set to 8. If a sample had more than 8 views, it was split into multiple samples.

3.3. Evaluation metrics

For the image segmentation task, we adopt several commonly used metrics [12] for model evaluation, namely Intersection over Union (IoU), Dice Similarity Coefficient (DSC), and Sensitivity (Sen), calculated as follows:

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{12}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{13}$$

$$Sen = \frac{TP}{TP + FN} \tag{14}$$

where TP, TN, FP and FN are number of true positive, true negative, false positive and false negative pixels, respectively.

Statistical analysis with paired student's t-test is performed on all indices, and $p < 0.05$ indicates statistically significant difference.

4. Results

4.1. Results of data augmentation

Fig. 7 shows four examples of data augmentation based on 3D view interpolation, including two linear scars and two patchy scars. The reconstructed 3D models, and all the original images and the generated projected images for each sample are listed. It can be observed that the 3D models can clearly reflect the 3D information of these scars, including their spatial position, geometric shape, and texture attributes. Moreover, the generated images can show the scars from different angles and distances.

To show the effectiveness of the proposed data augmentation and semi-supervised learning, we compare our first-stage segmentation results with that obtained using original images, and that obtained using offline random transformation (ORT) method. ORT refers to performing random flipping and rotation on the original image and masks to generate more supervised data. The number of generated images are set to the same. Table 2 lists the results of comparative experiments on data augmentation methods. Compared to training using the original images, the proposed data augmentation improves the IoU, DSC, and Sen by 1.97%, 1.37% and 1.28% respectively for linear scars, and by 5.24%, 3.97% and 4.21% respectively for patchy scars. The improvement of average indices is more significant for patchy scars with small sample size. Moreover, for patch scars, the standard deviation values are more significantly reduced. This indicates that, with the proposed data augmentation, the network performs consistently well for all samples across different folds. Therefore, the comparisons in both mean and standard deviation indicate improved generalization, especially for the dataset with smaller size. Although ORT can improve the performance of the network to a certain extent, it does not perform well on all folds of the patchy scars, still resulting in quite large variances. Statistical tests also show significant difference between indices of the proposed method and the other two, except for only one case indicated by "*" in Table 2.

4.2. Ablation study of MVCSNet

Ablation studies were conducted to show that the designs of MVCSNet all benefit the segmentation performance. For the experiments in this subsection, the output of ResUNet [32] with the proposed data augmentation method was used as the segmentation prior.

We first performed ablation experiments on the backbone of MVCSNet. Results with the ResNet18, the original ResNet34, ResNet34 without the first max pooling layer, ResNet34 without the fourth residual block, and the proposed modified ResNet34 without the first max pooling layer and fourth residual block, are listed in Table 3. For different backbones, the feature maps were interpolated to the same spatial size before going into the MVFI module. The computational complexity of these variations of MVCSNet is also compared in terms of the total number of parameters and floating point operations (FLOPs). In comparison, with the proposed modified ResNet34 as backbone, the indices are the highest in all cases with statistical tests showing significant difference. Larger differences are observed for linear scars, indicating that higher resolution contributes more to detection of thin structures. For model complexity, removing the maxpooling layer increases the resolution and therefore largely increases the FLOPs, while removing the fourth residual block reduces the number of parameters and slightly reduces the FLOPs. In general, the proposed modified ResNet34 achieves better segmentation results by preserving higher resolutions while maintaining decent computational complexity. In addition, using a lighter ResNet18 encoder can reduce the model complexity, but its indices are lower than both modified and original ResNet34. This may due to its weaker ability of feature extraction with less convolution layers.

Then, to verify the effectiveness of each component in MVCSNet, we conducted ablation experiments on the MVFI module, the segmentation prior, and multi-view co-segmentation. The results of these ablation experiments are listed in Table 4. The baseline in the first row is the ResUNet structure with the modified ResNet34 as the encoder. When prior segmentation is used without the MVFI module, it is multiplied to features of the lowest level. When multi-view segmentation is applied without the MVFI module, all view images of the same sample are input in a batch and processed together. When multi-view segmentation is not applied, images are randomly grouped into batches.

Comparing the first and second rows or the third and fourth rows, it can be seen that the prior segmentation has a significant impact on network performance. For linear scars, DSC is increased by 1.75% for single-view segmentation and 3.91% for multi-view co-segmentation. For patchy scars, due to the small amount of data, bigger difference can be observed. After adding prior segmentation, DSC is increased by 3.83% for single-view segmentation and 16.2% for multi-view co-segmentation. Comparing the first and third rows, both without the MVFI module, multi-view co-segmentation results in much worse performance than single-view segmentation. This is because during multi-view training, multi-view data is packed into samples. Therefore, the total number of input samples become smaller, and the diversity of data is far less than that of single-view training. Comparing the second and fourth rows, after adding the priors but still without the MVFI module, multi-view co-segmentation results in comparable performance with single-view segmentation. This shows that the multi-view co-segmentation without exploring the consistency information between

Fig. 7. Examples of data augmentation based on 3D interpolation projection. The first column is the reconstructed 3D models, images in red boxes are the original ones, and images in green boxes are the generated images.

Table 2
Results of experiments on data augmentation.

Methods	Linear scar			Patchy scar		
	IoU (%)	DSC (%)	Sen (%)	IoU (%)	DSC (%)	Sen (%)
No augmentation	79.76 ± 2.52	88.32 ± 1.77	88.45 ± 1.46	78.09 ± 11.82	86.58 ± 8.74	88.31 ± 7.93
ORT	80.67 ± 2.18	88.96 ± 1.49	88.73 ± 1.88	80.12 ± 10.22	88.36 ± 6.85	90.63 ± 4.79*
Proposed	**81.73 ± 2.30**	**89.69 ± 1.47**	**89.73 ± 1.79**	**83.33 ± 3.96**	**90.55 ± 2.62**	**92.52 ± 0.37**

Except the indices marked with *, all indices of "No augmentation" and "ORT" have statistically significant difference with $p < 0.05$, compared with the proposed method.

Table 3
Results of ablation experiments on backbones of MVCSNet.

Backbone	Linear scar			Patchy scar			Param	FLOPs
	IoU (%)	Dice (%)	Sen (%)	IoU (%)	Dice (%)	Sen (%)	(M)	(G)
Res18	79.94 ± 2.17	88.51 ± 1.55	88.68 ± 1.91	87.32 ± 1.70	92.92 ± 1.15	94.53 ± 0.27	13.49	51.50
Res34	80.32 ± 1.48	88.84 ± 0.96	89.18 ± 0.90	87.38 ± 1.71	92.95 ± 1.15	94.27 ± 0.67	23.60	70.88
Res34-M1	80.74 ± 1.15	89.10 ± 0.73	90.24 ± 0.83	87.32 ± 1.63	92.92 ± 1.10	94.09 ± 0.78	23.60	184.19
Res34-M2	81.77 ± 1.46	89.73 ± 0.94	90.39 ± 1.44	87.47 ± 1.77	92.99 ± 1.19	94.36 ± 0.45	9.30	63.55
Res34-M	**83.07 ± 1.60**	**90.51 ± 1.01**	**91.86 ± 0.82**	**87.52 ± 1.78**	**93.03 ± 1.20**	**94.68 ± 0.20**	9.30	154.90

Res34-M1: modified ResNet34, without the first max pooling layer.
Res34-M2: modified ResNet34, without the fourth residual block.
Res34-M: modified ResNet34 as proposed, without both the first max pooling layer and the fourth residual block.
All indices have statistically significant difference with $p < 0.05$, compared with the proposed Res34-M.

Table 4
Results of ablation experiments on multi-view components of MVCSNet.

#	MVFI	Prior	Multi-view	Linear scar			Patchy scar		
				IoU (%)	DSC (%)	Sen (%)	IoU (%)	DSC (%)	Sen (%)
1				79.32 ± 2.26	88.01 ± 1.70	89.08 ± 1.00	76.96 ± 10.65	86.25 ± 7.09	91.45 ± 2.14
2		✓		81.82 ± 2.03	89.76 ± 1.30	91.11 ± 1.23	82.55 ± 3.76	90.08 ± 2.49	92.92 ± 0.98
3			✓	76.33 ± 1.82	85.51 ± 1.43	85.61 ± 2.94	62.93 ± 13.95	74.22 ± 11.72	79.00 ± 10.13
4		✓	✓	81.34 ± 2.15	89.42 ± 1.37	89.96 ± 1.04	83.08 ± 3.96	90.42 ± 2.62	92.32 ± 0.36
5	✓	✓		81.52 ± 1.86	89.56 ± 1.21	90.15 ± 1.36	82.72 ± 3.68	90.19 ± 2.45	92.74 ± 0.11
6	✓	✓	✓	**83.07 ± 1.60**	**90.51 ± 1.01**	**91.86 ± 0.82**	**87.52 ± 1.78**	**93.03 ± 1.20**	**94.68 ± 0.20**

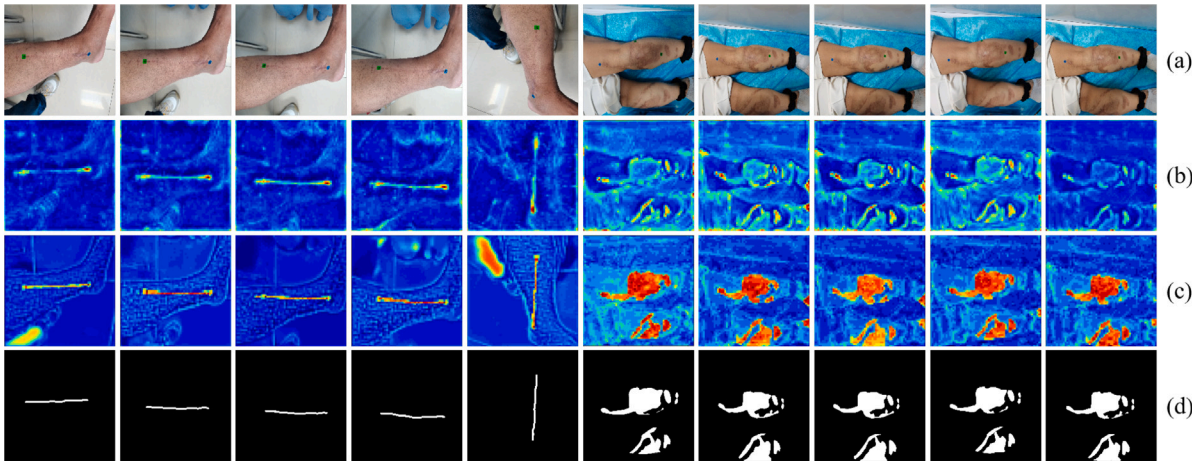

Fig. 8. Visual comparison of the feature maps of scars. (a) multi-view images (b) feature maps without MVFI module (c) feature maps with MVFI module (d) ground truth.

Table 5
p values of ablation experiments in Table 3.

	Linear scar			Patchy scar		
	IoU (%)	DSC (%)	Sen (%)	IoU (%)	DSC (%)	Sen (%)
1 vs. 2	3.34E−05	1.16E−05	7.29E−04	1.53E−04	1.49E−05	1.88E−03
1 vs. 3	3.11E−04	2.71E−03	5.65E−02*	1.29E−16	9.94E−15	1.31E−08
2 vs. 4	7.55E−04	6.60E−04	1.04E−03	5.59E−04	4.98E−03	1.22E−01*
2 vs. 5	9.70E−03	8.35E−03	4.78E−05	8.17E−03	1.81E−02	3.82E−02
3 vs. 4	1.06E−12	1.11E−08	3.08E−04	4.75E−09	5.13E−11	7.11E−10
4 vs. 6	2.92E−05	2.79E−06	7.56E−04	3.60E−07	4.30E−07	9.05E−03
5 vs. 6	5.29E−05	6.64E−04	2.33E−04	2.43E−09	2.98E−09	3.30E−03

* indicates no statistically significant difference with $p > 0.05$.

views, even with the help of prior segmentation, cannot improve the performance of the network. Comparing the second and fifth rows, for single-view segmentation, adding the MVFI module results in comparable performance. That means feature consistency optimization across samples is not effective. Comparing the fourth and the sixth rows, the effectiveness of the MVFI module for multi-view co-segmentation is demonstrated. Comparing the fifth and the sixth rows, the results of MVCSNet trained with multi-view data is the best. Compared with the results of first-stage segmentation (last row of Table 2), for linear scars, after using MVCSNet for second-stage optimization, the IOU, DSC, and Sen indices are improved by 1.34%, 0.82%, and 2.13% respectively. For patchy scars, the IOU, DSC, and Sen indices are improved by 4.19%, 2.48%, and 2.16% respectively.

The p-values of statistical tests for the results in Table 4 are listed in Table 5. There is statistically significant difference for all cases but two. This further proves the effectiveness of all proposed components in MVCSNet.

Fig. 8 presents the visual comparison between the output features of MVFI module and those without MVFI module. The features input into the decoder (of size 64×64) are compared. As can be seen from Fig. 8, the MVFI module effectively enhances the feature regions containing the scars, while suppressing background regions.

Fig. 9 shows the segmentation results of different model variations in the ablation experiments in Table 4. It can be seen that MVCSNet has the best segmentation results for scars after adding the proposed MVFI module, prior segmentation, and multi-view co-segmentation. Meanwhile, it can be found that solely incorporating multi-view co-segmentation or prior segmentation cannot effectively improve the segmentation results, and sometimes even causes the model to learn wrong information, thereby reducing the performance of the model.

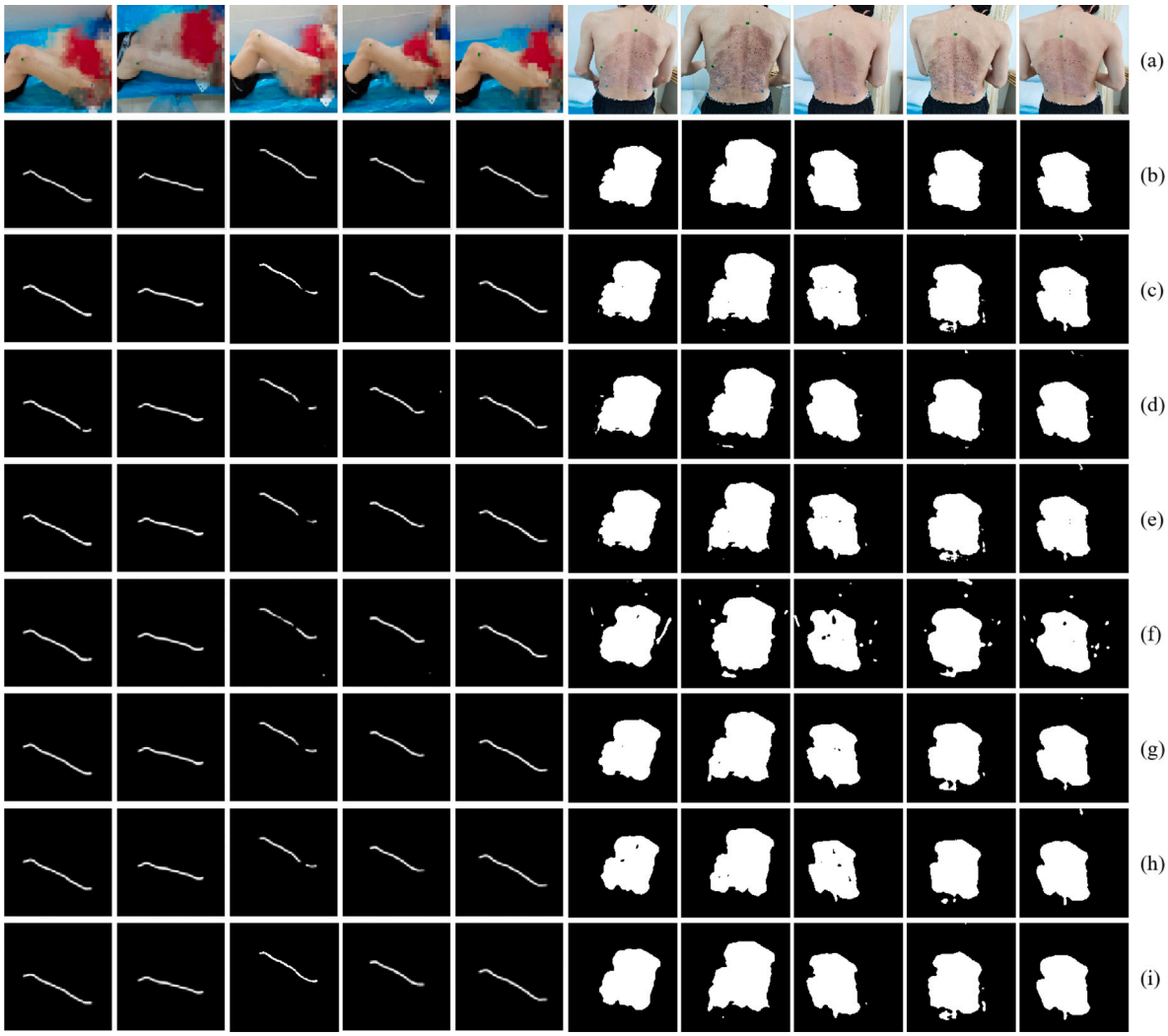

Fig. 9. Visual comparison of ablation experiments on MVCSNet. (a) multi-view images (b) ground truth (c) prior segmentation (d) single-view segmentation without MVFI (e) single-view segmentation with prior but without MVFI (f) multi-view co-segmentation without MVFI (g) multi-view co-segmentation with prior but without MVFI (h) single-view segmentation with MVFI (i) multi-view co-segmentation with MVFI.

Table 6
Results of comparative experiments.

Methods	Linear scar			Patchy scar			Param	FLOPs
	IoU (%)	Dice (%)	Sen (%)	IoU (%)	Dice (%)	Sen (%)	(M)	(G)
UNet [18]	73.65 ± 2.84	83.68 ± 2.28	83.54 ± 3.22	72.31 ± 12.43	82.90 ± 9.15	87.20 ± 3.64	8.64	131.63
SegNet [38]	76.48 ± 1.87	86.25 ± 1.42	87.98 ± 1.74	73.56 ± 10.35	83.30 ± 8.39	85.56 ± 7.56	38.44	75.45
FCN [39]	78.00 ± 2.35	87.10 ± 1.82	87.61 ± 2.06	76.41 ± 8.59	85.92 ± 6.01	89.17 ± 4.34	25.21	82.07
CENet [19]	79.13 ± 2.69	87.91 ± 1.81	89.15 ± 1.95	75.78 ± 14.20	85.09 ± 10.34	91.45 ± 4.21	29.01	71.21
DeepLabv3+ [40]	79.26 ± 2.74	88.06 ± 2.14	89.05 ± 2.27	74.56 ± 14.41	83.73 ± 11.60	86.30 ± 9.47	26.71	109.25
CPFNet [20]	78.80 ± 2.66	87.72 ± 1.84	88.40 ± 1.62	73.04 ± 15.10	82.53 ± 12.28	87.99 ± 8.02	30.65	64.58
PSPNet [41]	77.36 ± 2.65	86.80 ± 1.84	88.22 ± 1.58	73.11 ± 10.65	82.88 ± 9.19	85.45 ± 7.76	27.50	47.07
ResUNet [32]	79.76 ± 2.52	88.32 ± 1.77	88.45 ± 1.46	78.09 ± 11.82	86.58 ± 8.74	88.31 ± 7.93*	21.74	63.85
nnUNet [42]	70.14 ± 2.93	80.72 ± 2.30	85.80 ± 1.85	69.80 ± 9.10	81.09 ± 6.60	87.40 ± 5.58	33.48	460.79
Swin-Unet [43]	42.33 ± 6.15	55.71 ± 6.29	58.11 ± 5.96	62.94 ± 6.46	76.07 ± 4.74	88.47 ± 8.12	27.15	30.87
FANet [12]	80.26 ± 2.46	88.73 ± 1.63	89.21 ± 1.57	76.07 ± 13.29	85.43 ± 9.46	90.03 ± 4.91	22.65	64.27
SAAB [28]	80.49 ± 2.59	88.78 ± 1.74	89.01 ± 1.66	77.64 ± 12.58	86.03 ± 9.74	87.78 ± 7.58*	25.94	63.86
DOCS [29]	80.33 ± 2.96	88.66 ± 2.03	88.82 ± 2.31	78.37 ± 11.48	86.73 ± 8.59	89.53 ± 6.34	22.49	64.23
COSNet [30]	80.28 ± 2.98	88.67 ± 2.04	89.27 ± 2.47	77.35 ± 12.12	86.39 ± 8.43	90.62 ± 5.36	31.44	66.33
First stage	81.73 ± 2.30	89.69 ± 1.47	89.73 ± 1.79	83.33 ± 3.96	90.55 ± 2.62	92.52 ± 0.37	21.74	63.85
MVCSNet	**83.07 ± 1.60**	**90.51 ± 1.01**	**91.86 ± 0.82**	**87.52 ± 1.78**	**93.03 ± 1.20**	**94.68 ± 0.20**	31.04	218.74

Except the indices marked with *, all indices of other methods have statistically significant difference with $p < 0.05$, compared with the proposed MVCSNet.

4.3. Comparison with other methods

We compared the results of MVCSNet with other excellent deep learning-based single image segmentation algorithms. These networks are all encoder–decoder structures. For fair comparison, except for U-Net [18], nnUNet [42] and SwinUNet [43], we replaced the feature encoders of these networks with ResNet34 [33]. In addition, we also compared with three deep learning-based co-segmentation algorithms,

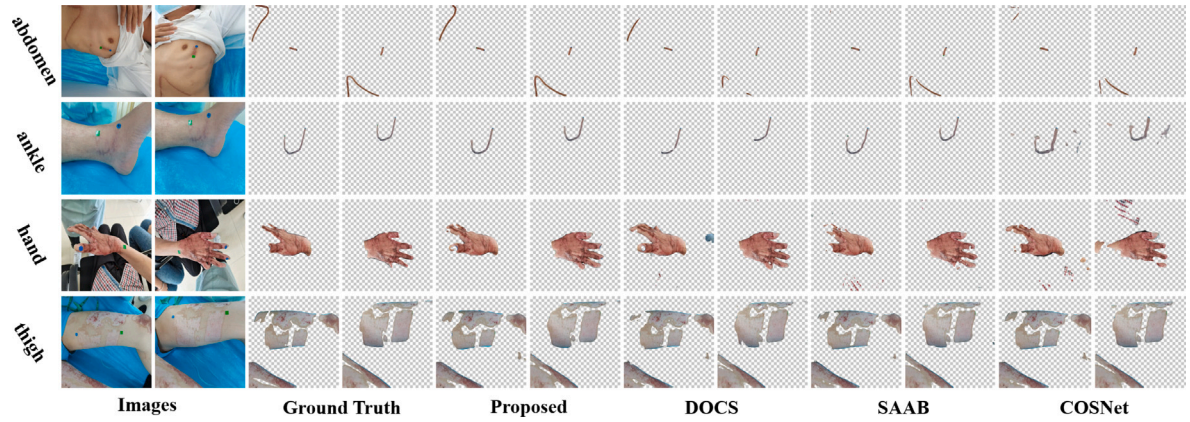

Fig. 10. Some qualitative comparison results generated by the proposed MVCSNet, DOCS and SAAB for co-segmenting scars from different parts of the human body.

SAAB [28], DOCS [29] and COSNet [30]. All of them require a pair of images as inputs. In our experiments, we fed two views of the same scar sample into these co-segmentation networks at a time.

Table 6 list the comparison results on the linear and patchy datasets. It can be observed that the proposed MVCSNet outperforms the other methods in terms of all metrics on both datasets. Especially, for the patchy scar dataset which has a smaller amount of data and is more challenging to segment, the IoU, DSC and Sen reach 87.52 ± 1.78%, 93.03 ± 1.20% and 94.68 ± 0.20%, respectively. For linear scars, the DSC of MVCSNet is at least 1.73% higher than other methods, and for patchy scars, that is 6.3%. Statistical tests also show significant difference in almost all cases. The outstanding results can be attributed to two main factors. Firstly, the data augmentation technique significantly contributes by enhancing the generalization capabilities of the ResUNet model, thereby improving the reliability of initial segmentation. Secondly, the MVFI module can effectively capture scar features from different views, and enhances regions of features with high similarity while suppressing those with low similarity through feature consistency optimization.

The last two columns of Table 6 give the number of network parameters and FLOPs of all methods, which are commonly used to measure the complexity of a deep learning model. Since the proposed MVCSNet uses ResUNet as prior segmentation, the numbers here are the sum of ResUNet and MVCSNet. The total parameters of the proposed method is not high, but the amount of calculation is large. This is due to the removal of the first maxpooling layer of ResNet34, which keeps the feature maps at a high resolution, and due to the cross-optimization of features from all views.

Fig. 10 shows the co-segmentation results by the proposed MVCSNet, compared with DOCS [29], SAAB [28] and COSNet [30] for two linear scars and two patchy scars from different parts of the human body. It can be observed that DOCS and SAAB often obtain more false negatives for linear scars, and more false positives for patchy scars, and COSNet results in more false negatives in some cases and more false positives in others. It can also be observed from the abdomen scar example that DOCS, SAAB and COSNet have poor recognition capabilities for objects with large view point changes, while MVCSNet is more robust to the change of viewpoints.

5. Conclusions and discussions

In this paper, a two-stage scar segmentation method based on convolutional neural networks is proposed. The multi-view information is utilized to improve the model performance in each stage, but from different perspectives. In the first stage, we propose a data augmentation strategy based on 3D view interpolation. Multi-view images are used collaboratively to reconstruct a 3D model, from which images of new viewpoints are simulated. These images improve the diversity of data for deep network training, prevent model overfitting and enhance the generalization ability. In the second stage, we propose the MVCSNet for multi-view co-segmentation. In MVCSNet, we design the MVFI module to capture semantic features and achieve multi-view feature interaction. In the MVFI module, the masked average pooling operation is applied to obtain the semantic descriptor of each view. Then the similarity matrices between each semantic descriptor and multi-view features are calculated and used for semantic level feature enhancement.

Experiments are performed on two datasets, one with linear scars and the other with patchy scars. The latter has less samples and greater variations in shape, size and texture, and thus is more challenging. The proposed 3D view interpolation obtains data with more diversity than simple flipping and rotation. The comparative experiments show the effectiveness of the proposed data augmentation method, especially for patchy scars with a small number of samples, greatly improving the generalization of the network. Ablation experiments on the MVCSNet prove the effectiveness of the MVFI module, the utilization of prior segmentation, and the multi-view co-segmentation strategy, and show that these components collaborate with each other and each of them is indispensable. In comparative experiments, we compared MVCSNet with some existing image segmentation and co-segmentation networks. The results show that the proposed method outperforms other related algorithms in all indices. Furthermore, unlike the compared co-segmentation algorithms which can only take two views, the MVCSNet can adapt to the input of any number of views, and can more accurately capture semantic features due to the introduction of prior segmentation result.

The proposed method offers support for image-based automatic 3D quantitative analysis of scars. Due to the three-dimensional nature of scars on the curved surface of human skin, conventional single-view analysis methods are inadequate for capturing accurate 3D information, especially for scars that span body parts, resulting in inaccurate measurements. Multi-view stereo (MVS) techniques reconstruct the 3D model of a scar and perform measurement in 3D space. In such settings, the proposed methods can fully exploit the information contained in multi-view images and locate the scar in each view more accurately and consistently. In the future, we will further study the 3D quantitative analysis of scars based on multi-view images. With multi-view images as input, quantitative measurements of scars, such as length, area, tortuosity, will be obtained automatically.

Fundings

This study was supported by grants from the National Key R&D Program of China (2022YFC3302001), the National Natural Science Foundation of China (62271337) and the National Key R&D Program of China (2018YFA0701700).

CRediT authorship contribution statement

Jian Zhou: Writing – original draft, Software, Methodology. **Yuqing Dai:** Visualization, Software. **Dongmei Liu:** Validation, Data curation. **Weifang Zhu:** Writing – review & editing. **Dehui Xiang:** Formal analysis. **Xinjian Chen:** Project administration, Funding acquisition. **Fei Shi:** Writing – review & editing, Supervision, Methodology. **Wentao Xia:** Writing – review & editing, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

[1] F.A. Bianchi, F. Roccia, P. Fiorini, S. Berrone, Use of patient and observer scar assessment scale for evaluation of facial scars treated with self-drying silicone gel, J. Craniofacial Surg. 21 (3) (2010) 719–723.

[2] P.D. Deps, H.P. Aborghetti, T.L. Zambon, V.C. Costa, J.D. dos Santos, S.M. Collin, P. Charlier, Assessing signs of torture: A review of clinical forensic dermatology, J. Am. Acad. Dermatol. 87 (2) (2022) 375–380.

[3] M.J. Flies, P.K. Larsen, N. Lynnerup, C. Villa, Forensic 3D documentation of skin injuries using photogrammetry: photographs vs video and manual vs automatic measurements, Int. J. Legal Med. 133 (3) (2019) 963–971.

[4] J. Jin, H. Li, Z. Chen, J. Sheng, T. Liu, B. Ma, S. Zhu, Z. Xia, 3-D wound scanner: A novel, effective, reliable, and convenient tool for measuring scar area, Burns 44 (8) (2018) 1930–1939.

[5] A.M. Elrefaie, R.M. Salem, M.H. Faheem, High-resolution ultrasound for keloids and hypertrophic scar assessment, Lasers Med. Sci. 35 (2020) 379–385.

[6] S. Maher, L. Dorko, S. Saliga, Linear scar reduction using silicone gel sheets in individuals with normal healing, J. Wound Care 21 (12) (2012) 602–609.

[7] M.F.A. Fauzi, I. Khansa, K. Catignani, G. Gordillo, C.K. Sen, M.N. Gurcan, Computerized segmentation and measurement of chronic wound images, Comput. Biol. Med. 60 (2015) 74–85.

[8] D.M. Dhane, M. Maity, T. Mungle, C. Bar, A. Achar, M. Kolekar, Fuzzy spectral clustering for automated delineation of chronic wound region using digital images, Comput. Biol. Med. 89 (2017) 551–560.

[9] L. Wang, P.C. Pedersen, D.M. Strong, B. Tulu, E. Agu, R. Ignotz, Q. He, An automatic assessment system of diabetic foot ulcers based on wound area determination, color segmentation, and healing score evaluation, J. Diabetes Sci. Technol. 10 (2) (2015) 421–428.

[10] S. Sarp, M. Kuzlu, M. Pipattanasomporn, O. Guler, Simultaneous wound border segmentation and tissue classification using a conditional generative adversarial network, J. Eng. 2021 (3) (2021) 125–134.

[11] C. Wang, D.M. Anisuzzaman, V. Williamson, M.K. Dhar, B. Rostami, J. Niezgoda, S. opalakrishnan, Z. Yu, Fully automatic wound segmentation with deep convolutional neural networks, Sci. Rep. 10 (1) (2020) 21897.

[12] P. Zhang, X. Chen, Z. Yin, X. Zhou, Q. Jiang, W. Zhu, D. Xiang, Y. Tang, F. Shi, Interactive skin wound segmentation based on feature augment networks, IEEE J. Biomed. Health Inf. 27 (7) (2023) 3467–3477.

[13] H. Wannous, Y. Lucas, S. Treuillet, Enhanced assessment of the wound-healing process by accurate multiview tissue classification, IEEE Trans. Med. Imaging 30 (2) (2011) 315–326.

[14] C. Liu, X. Fan, Z. Guo, Z. Mo, E.I.-C. Chang, Y. Xu, Wound area measurement with 3D transformation and smartphone images, BMC Bioinformatics 20 (2019) 724.

[15] R. Niri, E. Gutierrez, H. Douzi, Y. Lucas, S. Treuillet, B. Castañeda, I. Hernandez, Multi-view data augmentation to improve wound segmentation on 3D surface model by deep learning, IEEE Access 9 (2021) 157628–157638.

[16] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[17] M.H. Hesamian, W. Jia, X. He, P. Kennedy, Deep learning techniques for medical image segmentation: achievements and challenges, J. Digit. Imaging 32 (2019) 582–596.

[18] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: 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.

[19] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, J. Liu, CE-Net: Context encoder network for 2D medical image segmentation, IEEE Trans. Med. Imaging 38 (10) (2019) 2281–2292.

[20] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, X. Chen, CPFNet: Context pyramid fusion network for medical image segmentation, IEEE Trans. Med. Imaging 39 (10) (2020) 3008–3018.

[21] J. Cho, K. Lee, E. Shin, G. Choy, S. Do, How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? 2015, arXiv preprint arXiv:1511.06348.

[22] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 1–48.

[23] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, A. Haworth, A review of medical image data augmentation techniques for deep learning applications, J. Med. Imaging Radiat. Oncol. 65 (5) (2021) 545–563.

[24] Q. Xie, Z. Dai, E. Hovy, T. Luong, Q. Le, Unsupervised data augmentation for consistency training, Adv. Neural Inf. Process. Syst. 33 (2020) 6256–6268.

[25] Q. Xie, M.-T. Luong, E. Hovy, Q.V. Le, Self-training with noisy student improves imagenet classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10687–10698.

[26] L. Yang, W. Zhuo, L. Qi, Y. Shi, Y. Gao, St++: Make self-training work better for semi-supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4268–4277.

[27] S. Chen, G. Bortsova, A. García-Uceda Juárez, G. Van Tulder, M. De Bruijne, Multi-task attention-based semi-supervised learning for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22, Springer, 2019, pp. 457–465.

[28] H. Chen, Y. Huang, H. Nakayama, Semantic aware attention based deep object co-segmentation, in: Asian Conference on Computer Vision, Springer, 2018, pp. 435–450.

[29] W. Li, O.H. Jafari, C. Rother, Deep object co-segmentation, in: Asian Conference on Computer Vision, Springer, 2018, pp. 638–653.

[30] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, F. Porikli, See more, know more: Unsupervised video object segmentation with co-attention siamese networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3623–3632.

[31] J.L. Schonberger, J.-M. Frahm, Structure-from-motion revisited, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4104–4113.

[32] Z. Zhang, Q. Liu, Y. Wang, Road extraction by deep residual U-Net, IEEE Geosci. Remote Sens. Lett. 15 (5) (2018) 749–753.

[33] 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, pp. 770–778.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[35] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, Labelme: A database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (13) (2008) 157–173.

[36] Colmap open source library, https://github.com/colmap/colmap.

[37] OpenMVS: open Multi-View Stereo reconstruction library, https://github.com/cdcseacave/openMVS.

[38] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.

[39] 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, pp. 3431–3440.

[40] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, 2017, arXiv preprint arXiv:1706.05587.

[41] 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, pp. 2881–2890.

[42] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, NnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2) (2021) 203–211.

[43] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-Unet: Unet-like pure transformer for medical image segmentation, in: Proceedings of the European Conference on Computer Vision Workshops, ECCVW, 2022.