BMC Medical Imaging

RESEARCH Open Access

Auto-segmentation of surgical clips for target volume delineation in post-lumpectomy breast cancer radiotherapy

Xin Xie[1†], Peng Huang[2†], Zhihui Hu[2†], Yuhan Fan[2†], Jiawen Shang[2†], Ke Zhang[2*] and Hui Yan[2*]

Abstract

Purpose To develop an automatic segmentation model for surgical marks, titanium clips, in target volume delineation of breast cancer radiotherapy after lumpectomy.

Methods A two-stage deep-learning model is used to segment the titanium clips from CT image. The first network, Location Net, is designed to search the region containing all clips from CT. Then the second network, Segmentation Net, is designed to search the locations of clips from the previously detected region. Ablation studies are performed to evaluate the impact of various inputs for both networks. The two-stage deep-learning model is also compared with the other existing deep-learning methods including U-Net, V-Net and UNETR. The segmentation accuracy of these models is evaluated by three metrics: Dice Similarity Coefficient (DSC), 95% Hausdorff Distance (HD95), and Average Surface Distance (ASD).

Results The DSC, HD95 and ASD of the two-stage model are 0.844, 2.008 mm and 0.333 mm, while their values are 0.681, 2.494 mm and 0.785 mm for U-Net, 0.767, 2.331 mm and 0.497 mm for V-Net, 0.714, 2.660 mm and 0.772 mm for UNETR. The proposed 2-stage model achieved the best performance among the four models.

Conclusion With the two-stage searching strategy the accuracy to detect titanium clips can be improved comparing to those existing deep-learning models with one-stage searching strategy. The proposed segmentation model can facilitate the delineation of tumor bed and subsequent target volume for breast cancer radiotherapy after lumpectomy.

Keywords Segmentation, Surgical clip, Target volume, Deep learning, Radiotherapy

†Xin Xie, Peng Huang, Zhihui Hu, Yuhan Fan and Jiawen Shang contributed equally to this work and co-first authors.

*Correspondence:
Ke Zhang
zk110105@163.com
Hui Yan
hui.yan@cicams.ac.cn
[1]Department of Radiation Oncology, Clinical Oncology School of Fujian Medical University, Fujian Cancer Hospital, Fuzhou 350014, China
[2]Department of Radiation Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China

Introduction

Breast-conserving surgery (BCS) and whole-breast irradiation (WBI) is a standard alternative to mastectomy for most patients with early breast cancer [1, 2]. Boosting the tumor bed in addition to WBI can improve local control with mild side effects and acceptable cosmetic outcome [3, 4]. Sequential boost to the tumor bed (TB) was frequently used but would extend treatment courses. Delivering a concurrent boost dose, simultaneously integrated boost (SIB), to TB could result in greater convenience for patients and utilization of radiation [5, 6]. SIB provides a localized dose enhancement in the area at highest risk without prolonging treatment duration. However, to reduce radiation toxicities to surrounding the organs at risks (OARs), such as lung, the advanced techniques for target volume delineation, treatment planning and delivery, and quality control are required [7, 8].

For many years, TB is defined according to a combination of information: tumor mass, surgical clip, seroma, scar, tumor cavity, etc [9]. According to the recommendations of the International Commission on Radiation Units (ICRU) report 83, it is delineated based on the surgical clip, the residual seroma, and the tumor cavity on post-operative CT image [10]. Studies have shown that the inter-operator variability of TB contouring is large and geographical misses are frequent [11]. Titanium clips are recommended to delineate the lumpectomy cavity region more accurately as anatomy on post-operative CT may be quite different from the one on pre-operative CT [12]. It is also essential for SIB as it uses smaller planning target volumes to reduce the risk of late normal tissue toxicity [13, 14].

The threshold-based methods are frequently used in clip segmentation as its intensity or Hounsfield Unit (HU) is much higher than that of soft tissue on CT image. Kazemimoghadam et al. utilized single-threshold-based method to segment clips with the mask of breast [15]. Buehler et al. applied top-hat transformation to correct for uneven background illumination after binarization of the CTs with threshold-based method [16]. Because the gray value of bony structures is higher than the threshold, this may cause wrong segmentation. Ng et al. made further step to take the radius and the size of the segmented regions into account [17]. However, the limitations of threshold method are apparent. Firstly, it is required to specify the search region for clips, such as the mask of breast and the location of clips in adjacent images. Secondly, the HU of bony structures and metal is close to that of clips and cause wrong segmentation of clips.

Over the past decade, there have been many convolutional neural networks (CNN) proposed to semantic segment 3D medical image such as Computed Tomography (CT) [18, 19] and Magnetic Resonance Imaging (MRI) [20, 21]. Through supervised learning [21, 22], The CNN can learn information about the target, including its surroundings and the global environment, in addition to its size and location. To adapt to 3D medical image segmentation, Cicek et al. [23] and Milletari et al. [24] employed 3D operators, such as 3D convolution kernel and 3D pooling kernel, in the U-Net [25] architecture. Isensee et al. [26] proposed nnU-Net which consists of U-Net and 3D U-Net. It outperformed in many challenges by enhancing data preprocess and advanced training strategy.

In the recent years, Transformers-based networks were developed in the field of natural language processing (NLP) [27, 28], and very powerful for tasks such as translation with attention mechanism. To introduce Transformers into the field of image processing, Dosovitskiy et al. proposed Vision Transformers to split the image into many sub-blocks to simulate the words in NLP [29]. Hatamizadeh et al. [30, 31] introduced Transformers in U-Net by replacing the encoder with transformer blocks. These transformer blocks were connected to the decoder with skip-connection structure. Chen et al. [32] proposed to split the feature map from encoder into sub-blocks, and applied Transformers blocks to these sub-blocks.

The most networks for medical image segmentation are specify for large organs or objects, such as liver and tumor mass [33]. For the titanium clips which are smaller in size, the existing segmentation networks should be adjusted to address the issue of smaller markers. To solve it, Gao et al. proposed FocusNetv2 to segment small organs from CT [33]. It divided the segment task into multiple stages corresponding to four networks. A large amount of annotated data is needed to train this model. Tao et al. proposed Spine-transformers to segment the vertebras in two networks [34]. Comparing to vertebras, the clip is very small and hardly segmented accurately.

In this paper, we propose a two-stage model to locate the potential clip region and then detect clip locations from post-operative CTs for post-lumpectomy breast radiotherapy. The box region of clips is first identified from CT by the Location Net, and then the location of clip is detected by the Segmentation Net. The rest of this paper is organized as follows. In Sect. "Materials and methods", the architecture of the proposed two-stage segmentation model is introduced. In Sect. "Results", the ablation studies with and without feature map are performed and analyzed. The proposed model is also compared with the other existing deep-learning segmentation models. In Sect. "Discussion", the merits and limitations of this study are discussed.

Materials and methods

Patient dataset

101 breast cancer patient undergone breast-conserving surgery (BCS) and eligible for whole breast irradiation

(WBI) plus boost irradiation were collected retrospectively in two hospitals, including 82 cases from Fujian Cancer Hospital and 19 cases from National Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College (CAMS). The median age of patients was 52 years (range, 42–60 years), and the pathological diagnosis was all invasive ductal carcinoma with a stage of T1-T2N0M0. All patients underwent a lumpectomy with sentinel lymph node dissection. Tumor-negative margins were ensured during a single operation. Equal or more than 5 titanium clips were used to mark the boundaries of the lumpectomy cavity. This study was approved by the Institutional Ethics Committee of Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College/Clinical Oncology School of Fujian Medical University, Fujian Cancer Hospital (The ethics approval number: K2023-345-01). Informed Consent was waived in this retrospective study.

Patient CTs were non-contrast-enhanced and acquired averagely 10 weeks after surgery and used for radiotherapy treatment planning. In postoperative CT simulation, the patients were in the supine position, immobilized on a breast bracket with no degree of incline, and placed using arm support (with both arms above the head). All CT images were scanned using a Somatom Definition AS 40 (Siemens Healthcare, Forchheim, Germany) or a Brilliance CT Big Bore (Philips Healthcare, Best, the Netherlands). Their dimensions are 512×512 with the slice number varied from 35 to 196. The slice thickness are 5.0 mm (some special cases are 3.0 mm). The pixel sizes of these CT images vary from 1.18 mm to 1.37 mm. All clip contours were delineated manually by the same physician and confirmed by one senior physician.

Two-stage segmentation model

A two-stage model is proposed for clip segmentation in post-lumpectomy breast cancer radiotherapy. As shown in Fig. 1, it consists of two components: Location Net and Segmentation Net. In the first stage, the Location Net is used to search for the region of interest (ROI) which contains all titanium clips. Its input is the CT images with HU rescaled to different window levels. In the second stage, the Segmentation Net is used to search for the location of clips. Its input is the cropped CT images containing all clips and the feature maps obtained from the Location Net. As shown in Fig. 1, a modified U-Net, Res-SE-U-Net, is used in both stages. It consists of the up-sampling path, down-sampling path, and 5 skip-connection structures, which can utilize the multi-scale features and relieve vanishing gradient problem. In addition Res-SE-U-Net has 11 Res-blocks proposed in ResNet [35] and 1 SE-blocks proposed in SENet [36]. Both Res-block

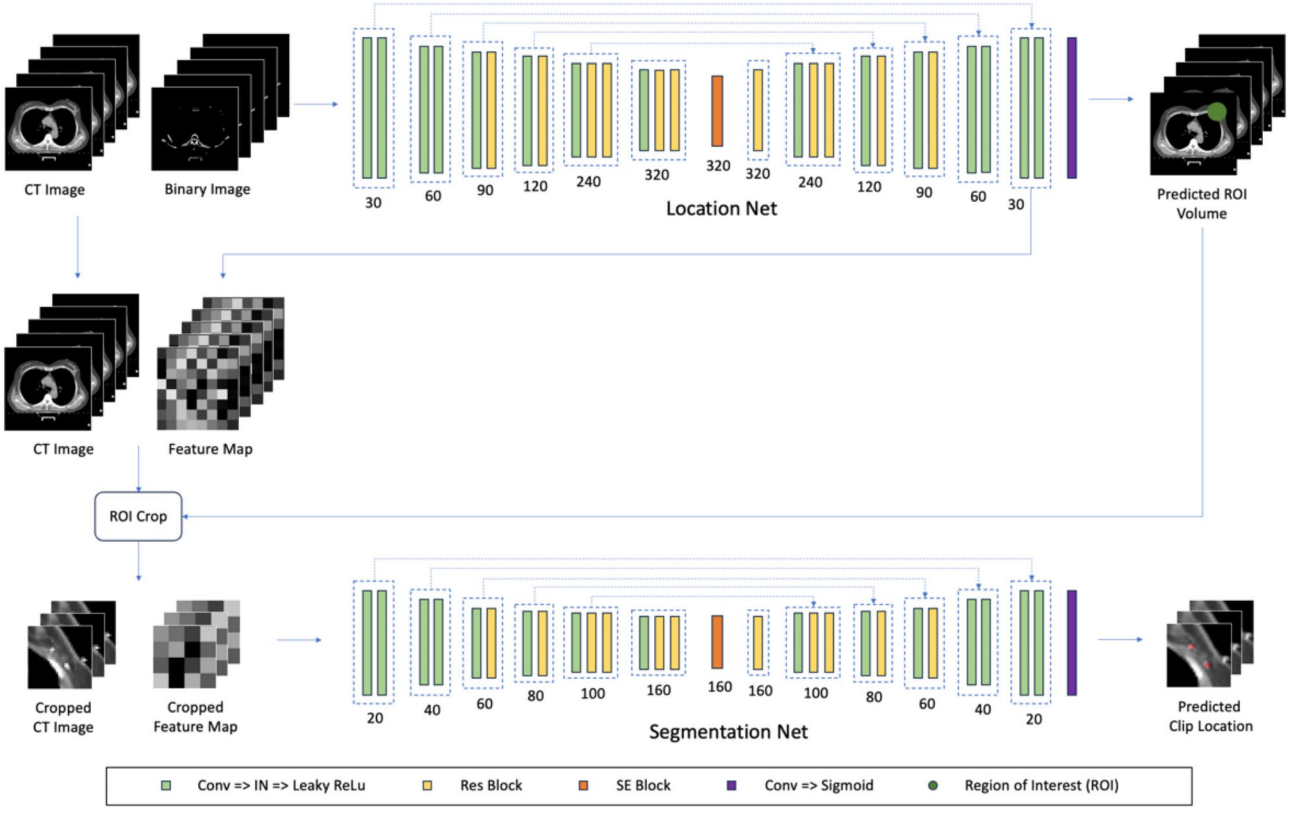

Fig. 1 The network architecture of the proposed model

and SE-block improve the network ability in extracting image features.

The Location Net of the first stage is a Res-SE-U-Net with two 3D input channels and one 3D output channel. The first input channel is the CT images with HU in [– 200, 200]. The second input channel is the binary images with HU of CT images set to 0 when its value less than 300 and 1 otherwise. This processing enhances the features of bony structures and metal objects on the input image. The output is the predicted label image with 1 for clip pixel and 0 for non-clip pixel. As the label images obtained, the coordinates of the center of all clip pixels is determined and used as the center of volume to crop CT images in next stage. In the model training, the random patches of foreground/background are sampled at ratio 1:1. In the model inference, a sliding window approach with an overlapping of 0.25 for neighboring voxels was used.

The Segmentation Net is another Res-SE-U-Net with two 3D input channels and one 3D output channel. The first input channel is the CT image with HU in [-200 200], which is the same as that of the Location Net. The second input channel is the feature maps which are copied from the last layer of the Location Net. This feature map contains condensed local and regional information of CT images. As the center of clips determined in previous stage, the ROI volume are cropped from the original CT and feature maps. As a result, the portions of CT images and feature maps in the dimensions of $96 \times 96 \times 96$ are obtained and fed into Res-SE-U-Net in the second stage. The output is the predicted label image with 1 for clip pixel and 0 for non-clip pixel. In the model training, a center cropping in the size of $96 \times 96 \times 96$ was applied to the input images based on the center of all clips in the labels with random spatial shifts. In the model inference, the sliding window approach with an overlapping of 0.25 is also applied on the cropped CT image and feature map.

Since the air and treatment couch in the background occupied most of space in CT images while the human body only takes a small portion of the images, it is helpful to remove their effects and focus more on the interested region such as human body. For this goal, the threshold method followed by the morphological method was applied to the CT images to enhance the pixels of human body. First, the air pixels in CT image are removed by setting all pixels with HU less than – 150 to 0. Next morphological opening and closure methods were applied to correct for uneven foreground. Last, the largest connected region was selected to remove the pixels of treatment couch from CT images. To maintain consistent resolution across all images, CT and the corresponding label images were resampled to $2 \times 2 \times 2$ mm in voxel sizes.

The datasets from two hospitals are randomly divided into training (90 cases), validation (6 cases), and test (5 cases) sets. Hierarchical sampling is applied based on the hospital source of the case. The training, validation, and test sets were used for model learning, hyper-parameter selection, and model evaluation, respectively. In the process of hyper-parameter selection, a grid search strategy was applied, and the network hyper-parameter including learning rate, weight of loss function, and number of net layers were adjusted based on the DICE of the validation set. During this process, the test set was not used to avoid data leakage.

The Location Net is first trained. During the training of the Segmentation Net, the parameters of the Location Net are fixed. Data augmentation strategies were used including random rotation, random flip in axial, sagittal and coronal views and random shift intensity in the range from 0.9 to 1.1. The model is implemented with PyTorch, Lightning and Monai on a single NVIDIA RTX 3090. The ADAM optimizer is used to train the models with a linear warmup of 50 epochs and using a cosine annealing learning rate scheduler. All models use a batch size of 2 for 1000 epochs, and initial learning rate of 1e-4.

Experiments

Ablation studies were conducted to evaluate the impact of feature map on the model performance. Four combinations are tested with the same network architecture but with different input images. In the first test the Location Net has one input channel (CT image) while the Segmentation Net has one input channel (CT image). In the second test the Location Net has two input channels (CT image and Binary image) while the Segmentation Net has one input channel (CT image). In the third test the Location Net has one input channel (CT image) while the Segmentation Net has two input channels (CT image and Feature Map). In the four test the Location Net has two inputs (CT image and Binary Image) while the Segmentation Net has two inputs (CT image and Feature Map).

The proposed method is also compared with three existing deep-learning models: 3D U-Net, V-Net and UNETR. All the models are trained and validated on the same CT datasets as the proposed model. 3D U-Net [23] and V-Net [24] are both modified U-Net and use 3D operators for 3D medical image segmentation. The architecture of 3D U-Net is similar to the architecture of U-Net which consists of an encoder with down-sampling and a decoder with an up-sampling. 3D U-Net replaces the 2D convolutional kernels in U-Net with 3D convolutional kernels and 2D pooling kernels with 3D pooling kernels. In addition, 3D U-Net introduces weighted softmax as its objective function to focus itself more on the specified target. Compared with U-Net learning from

each slice separately, 3D U-Net can utilize the 3D features between the slices of 3D medical images.

Compared with 3D U-Net, V-Net applied 3D convolutional kernels with a stride of 2, instead of 3D pooling kernels to down-sample the feature map in the decoder. V-Net also introduced residual connections from ResNet to relieve the vanishing gradient caused by the large scale of the network. In addition, a novel objective function based on the Dice coefficient is introduced to solve the problem that the foreground region with a small volume is often missing or only partially detected.

To apply Transformers in the field of 3D medical image segmentation, UNETR [30] replaces the encoder of U-Net with 12 Transformer modules and connected them to the decoder every 3 modules with skip connections directly. Since Transformer only works on 1D sequences, UNETR needs to convert the image into several sequences like sentences and words in NLP. Thus, UNETR split the image into patches without overlapping and then flattened them as sequences. As a result, the model can further utilize the global and local features to improve the performance by the attention mechanism from the Transformer.

Evaluations

To quantify the segmentation accuracy, three metrics are employed including DSC, 95% Hausdorff Distance (HD95), and Average Surface Distance (ASD). let G_i and P_i denote the ground truth and prediction values for voxel i and G' and P' denote ground truth and prediction surface point sets respectively.

DSC measures the spatial overlap between the predicted segmentation and the ground truth segmentation defined as follows:

$$Dice\,(G,\,P) = \frac{2\sum_i G_i P_i}{\sum_i G_i + \sum_i P_i} \tag{1}$$

The value of a DSC ranges from 0 to 1, while 0 indicates that there is no overlap between the predicted and ground truth segmentation, and 1 indicates that they overlap completely.

HD quantifies how closely the surfaces between the predicted and the ground truth segmentation. It measures the max distances between ground truth and prediction surface point sets defined as follows:

$$HD\,(G',\,P') = \max\left\{\max_{g'\in G'}\min_{p'\in P'}||g'-p\prime||,\max_{p'\in P'}\min_{g'\in G'}||p'-g\prime||\right\} \tag{2}$$

HD is sensitive to the edges of segmented regions. To eliminate the effects of outliers HD95 is mostly used and calculates the 95% largest distances for model evaluation.

ASD is also used to quantify the quality of segmentation result. It measures the average distance between the ground truth and prediction surfaces instead of the overlap of two volumes, and it is formally defined as follows:

$$ASD\,(G',\,P') = \frac{1}{|G'|+|P\prime|}\left(\sum_{p'\in P'}\min_{g'\in G'}||p'-g'||+\sum_{g'\in G'}\min_{p'\in P'}||g'-p'||\right) \tag{3}$$

The value range of both HD95 and ASD is greater than 0, while the larger value indicates the long distance between the surface of predicted and ground truth segmentation.

Results

Ablation studies

For model with location network, the segmentation accuracy of the proposed model with and without feature inputs is shown in Table 1. For model without both feature maps, the DSC is 0.722 while HD95 and ASD are 2.290 mm and 0.819 mm, respectively. For model with binary image and without feature map, the DSC increases by 3.8% while HD95 and ASD decrease by to 5.4% and 30.1%, respectively. For model without binary image and with feature map, the DSC increases by 13.5% while HD95 and ASD decrease by to 12.6% and 53.7%, respectively. For model with binary image and feature map, the DSC increases by 16.8% while HD95 and ASD decrease by to 12.3% and 59.3%, respectively. The result shows that the input of feature map has the greater impact than that of binary image on the segmentation accuracy. Without both of these feature inputs the segmentation accuracy of the proposed model could be significantly reduced.

For model without location network (only segmentation net is used), the DSC, HD95 and ASD are 0.714 ± 0.061, 2.372 ± 0.304 mm and 0.829 ± 0.148 mm, respectively. Comparing to the model with both location and segmentation nets, the DSC decreases by 1.1-15.4% while HD95 and ASD increase by 3.5-18.1% and 1.2-148.9%, respectively. The introduction of location network can provide valuable feature map around clips. If this feature map wasn't used in segmentation net, then the accuracy is improved slightly. If it wasn't used in segmentation net, the accuracy is improved apparently as

Table 1 Ablation studies for impact of the feature inputs

Location net		Segmentation net		Metrics		
CT Image	Binary Image	CT Image	Feature Map	DSC	HD95 (mm)	ASD (mm)
✓	×	✓	×	0.722 ± 0.077	2.290 ± 0.361	0.819 ± 0.139
✓	✓	✓	×	0.749 ± 0.206	2.166 ± 0.331	0.573 ± 0.608
✓	×	✓	✓	0.819 ± 0.047	2.000 ± 0.405	0.379 ± 0.102
✓	✓	✓	✓	0.844 ± 0.064	2.008 ± 0.016	0.333 ± 0.138

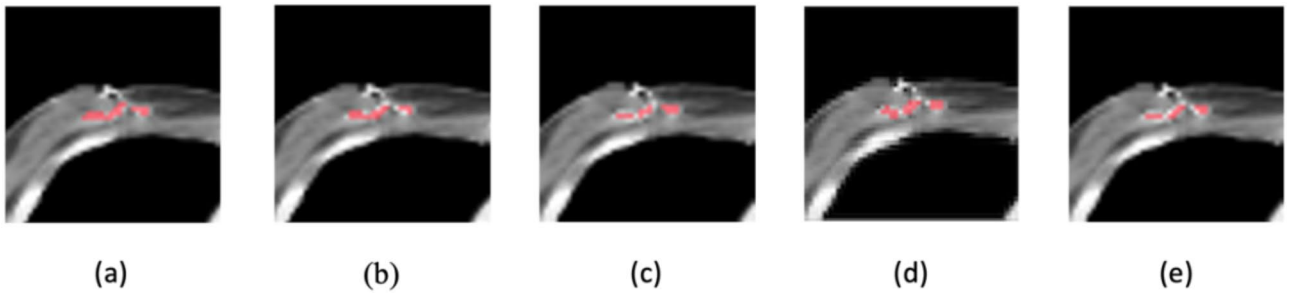

Fig. 2 Clips segmented by the proposed model with different feature inputs. (**a**) The model output without both binary image and feature map, (**b**) The model output with binary image and without feature map input, (**c**) The model output without binary image and with feature map input, (**d**) The model output with both binary image and feature map, and (**e**) True clip contour delineated manually by physician. The red areas indicate the regions detected by the proposed models or delineated by physician

Table 2 Comparison of the proposed and three deep-learning models

Models	DSC	HD95 (mm)	ASD (mm)
Threshold	0.403 ± 0.161	43.871 ± 35.208	12.352 ± 13.563
3D U-Net	0.681 ± 0.101	2.494 ± 0.304	0.785 ± 1.550
V-Net	0.767 ± 0.088	2.331 ± 0.175	0.497 ± 0.809
UNETR	0.714 ± 0.305	2.660 ± 1.165	0.772 ± 0.889
Proposed	0.844 ± 0.064	2.008 ± 0.016	0.333 ± 0.138

shown in third and fourth lines of Table 1. This result shows that the location net is important for the accuracy of clip segmentation.

The segmentation results of a patient case by the proposed model with different feature inputs are shown in Fig. 2a and d, and the clips delineated manually by physician is shown in Fig. 2e. In Fig. 2a and b two clips on the left and in the middle are detected as one clip by models. Both models in Fig. 2a and b have no input channel of feature map in the second stage. In Fig. 2c and d, three clips are clearly identified and there is no wrong segmentation. Both models in Fig. 2c and d have input channel of feature map in the second stage. Among the four models, the proposed model shows the closet result to that of manual result, and which outperforms the other three models.

Model comparison

Table 2 shows the segmentation accuracy by the proposed model, threshold method and three existing deep-learning models. For 3D U-Net, the DSC is 0.681 while HD95 and ASD are 2.494 mm and 0.785 mm, respectively. For V-Net, the DSC increases by 12.6% while HD95 and ASD decrease by to 6.5% and 36.7%, respectively. For UNETR, the DSC increases by 4.85% and ASD decrease by 1.65%, but HD95 increase by to 6.66%. For the proposed model with both feature inputs, the DSC increases by 23.9% while HD95 and ASD decrease by to 19.5% and 57.5%. Compared with deep learning-based methods, the threshold method reduces DICE by 50%, while HD95 and ASD have significantly increased due to misidentification by the threshold method. Among the five models, the proposed model has the highest DSC and lowest HD95 and ASD, which outperforms the other four models.

The segmentation results of a patient case by the proposed model and the other three existing deep-learning models are shown in Fig. 3a and d, and the clips delineated manually by physician is shown in Fig. 3e. In Fig. 3a there is a wrong clip labeled on the surface of breast by 3D U-Net. The shape of this object has similar density to that of clip but its location is on the surface which is wrong. In Fig. 3b the left portion of middle clip is labeled as a portion of the left clip, and the right portion of

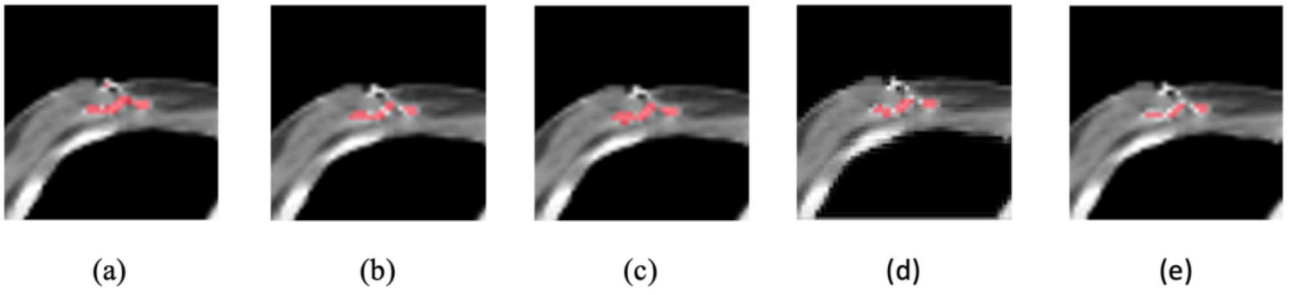

Fig. 3 Clips segmented by the proposed and three other existing deep-learning models. (**a**) The segmentation result by 3D U-Net, (**b**) The segmentation result by V-Net, (**c**) The segmentation result by UNETR, (**d**) The segmentation result by the propose method, and (**e**) Clip contour delineated manually by physician. The red areas indicate the regions detected by the proposed models or delineated by physician

middle clip is missed by V-Net. In Fig. 3c the left clip and the middle clip are segmented but labeled as one clip by UNETR. In Fig. 3d three clips are clearly segmented and correctly labeled by the proposed model. Among the four tested models, the proposed two-stage model outperforms the other three models and shows the closet result to that of manual one.

Discussion

In clinical practice, the surgical clip is the major indicator for the tumor bed. The physician needs to locate the clips and expand certain margin to form target volume of the boost region. With the automatic clip contouring method, the workload for the physician to manually locate the clips is greatly reduced. As surgical clip is small and adjacent to rib, to identify them from CT images is not easy. To build a deep-learning model on a training set consisting of paired CT image and label image is an efficient way. But, as the clip is too small, the focus of network learning could be missed or distracted by the larger object or textures.

To solve this issue, a two-stage model is proposed to deal with segmentation in two different resolutions. A Location Net is first applied to search the region of interest containing all clips, and then a Segmentation Net is used to search for accurate location of clips on the enlarged image (cropped ROI volume). As demonstrated in Fig. 3, the two-stage model can not only detect clips from soft tissue with higher contrast background, but also can distinguish the clip from the adjacent bony structures with lower contrast background.

The role of binary image and feature map for the proposed model is crucial. Binary Image provides the regions with higher density that the clips may exist, although these regions could contain bony structures and metal. With the introduction of binary image, DSC increases modestly while ASD decreases considerably as shown in Table 1. Feature map provides more useful features from the last layer of the Location Net. With the introduction of feature map, DSC increases considerably while ASD decreases considerably as shown in Table 1. While both binary image and feature map introduced, the DSC and ASD can be further improved as shown in Table 1. This indicates that the local and regional point and texture information could be utilized to improve the segmentation accuracy of the learning model.

Compared with previous studies that used threshold-based methods, the proposed method can automatically searches for the area where the clips are located, and require no need for beforehand assignment of the search area for the clips. Besides, the proposed method can reduce the inaccurate-segmentation caused by adjacent bone structures of which HU values are relatively close. Comparing with the other existing deep-learning models, the two-stage model achieved the better results in clip segmentation in this study. The proposed model detects clips correctly from the normal tissue and bony structures. However, the other models could fail due to the impact of surrounding artifacts. For example as shown in Fig. 3a the pixel on the surface of breast is wrongly detected as clip due the similar contrast of high-density object. Even clips are correctly segmented from normal tissues, their volume could be wrong. For example, as shown in Fig. 3b part of clip is detected and labeled as another clip on the left. In addition, the edge of clips could be merged with the neighboring high-density object. For example as shown in Fig. 3c two close clips are segmented but labeled as one clip. Therefore, the segmenting clip from normal tissue in CT image is a challenging task, which needs advanced deep-learning models and searching strategies.

Although the excellent performance of the proposed two-stage model for clips segmentation in this study, there are several aspects to be improved. First, there are few patient cases undergone breast-conserving surgery and eligible for whole breast irradiation plus boost irradiation. To build a robust model, more representative cases collected from multiple institutes is needed. In this study, only two institutes' patient data are available and the total number of cases is 101. This is insufficient in training a model with high generalizability and only used to demonstrate the effectiveness of the deep-learning model. In the future, it is advantageous to introduce pre-training model learned in the other similar medical image application, which can relieve the burden of the insufficient training data. Second, the proposed model employed two generic networks with the same architectures. It could be improved by adopting different networks which are more appropriate for the respective goals. In addition, it could be beneficial to introduce attention mechanism and adversarial network to further improve model performance.

Conclusion

The proposed two-stage model provides an effective way to segment clip for target volume delineation in post-lumpectomy breast cancer radiotherapy. The result shows that with the feature inputs the segmentation accuracy could be improved significantly. Also the propose model outperforms the other existing deep-learning models due to the introduction of two-stage searching strategy. It is promising to apply the current model to automatic delineation of target volume in integrated tumor bed boost of whole-breast irradiation after lumpectomy.

Abbreviations

CT	Computed tomography
DSC	Dice Similarity Coefficient
HD95	Hausdorff Distance

ASD	Average Surface Distance
BCS	Breast-conserving surgery
WBI	Whole-breast irradiation
SIB	Simultaneously integrated boost
OARs	Organs at risks
ICRU	The International Commission on Radiation Units
CNN	Convolutional neural networks
NLP	Natural language processing
ROI	The region of interest

Acknowledgements
Not applicable.

Author contributions
XX and PH performed the experiments and have made substantial contributions to the conception. ZH provided study materials or patients. YF and JS analyzed and interpreted the data. KZ collected the data. HY supervised the whole study. All authors wrote and have approved the manuscript.

Funding
This work was supported by the Joint Funds for the Innovation of Science and Technology, Fujian province (No.2023Y9442), and Fujian Provincial Health Technology Project (No. 2023GGA052), Fujian Clinical Research Center for Radiation and Therapy of Digestive, Respiratory and Genitourinary Malignancies (2021Y2014), the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (No. 2024-RW320-05), the National High Level Hospital Clinical Research Funding (No. 2022-CICAMS-80102022203).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The need for informed consent was waived because of the retrospective nature of the study. The Institutional Ethics Committee of Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College/Clinical Oncology School of Fujian Medical University, Fujian Cancer Hospital approved this study (The ethics approval number: K2023-345-01).

Consent for publication
The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests
The authors declare no competing interests.

Received: 9 March 2024 / Accepted: 13 March 2025
Published online: 21 March 2025

References

1. Dzhugashvili M, Veldeman L, Kirby AM. The role of the radiation therapy breast boost in the 2020s. Breast. 2023;69:299–305.
2. Smith BD, Bellon JR, Blitzblau R, Freedman G, Haffty B, Hahn C, Halberg F, Hoffman K, Horst K, Moran J, et al. Radiation therapy for the whole breast: executive summary of an American society for radiation oncology (ASTRO) evidence-based guideline. Practical Radiation Oncol. 2018;8(3):145–52.
3. Franco P, Cante D, Sciacero P, Girelli G, La Porta MR, Ricardi U. Tumor bed boost integration during whole breast radiotherapy: A review of the current evidence. Breast Care. 2014;10(1):44–9.
4. Shah C, Fleming-Hall E, Asha W. Update on accelerated whole breast irradiation. Clin Breast Cancer. 2023;23(3):237–40.
5. Bartelink H, Maingon P, Poortmans P, Weltens C, Fourquet A, Jager J, Schinagl D, Oei B, Rodenhuis C, Horiot J-C, et al. Whole-breast irradiation with or without a boost for patients treated with breast-conserving surgery for early breast cancer: 20-year follow-up of a randomised phase 3 trial. Lancet Oncol. 2015;16(1):47–56.
6. Yu T, Li Y, Sun T, Xu M, Wang W, Shao Q, Zhang Y, Li J, Yu J. A comparative study on hypofractionated whole-breast irradiation with sequential or simultaneous integrated boost on different positions after breast-conserving surgery. Sci Rep. 2021;11(1):18017.
7. Freedman GM, White JR, Arthur DW, Allen Li X, Vicini FA. Accelerated fractionation with a concurrent boost for early stage breast cancer. Radiother Oncol. 2013;106(1):15–20.
8. Wu S, Lai Y, He Z, Zhou Y, Chen S, Dai M, Zhou J, Lin Q, Chi F. Dosimetric comparison of the simultaneous integrated boost in Whole-Breast irradiation after Breast-Conserving surgery: IMRT, IMRT plus an electron boost and VMAT. PLoS ONE. 2015;10(3):e0120811.
9. Coles CE, Wilson CB, Cumming J, Benson JR, Forouhi P, Wilkinson JS, Jena R, Wishart GC. Titanium clip placement to allow accurate tumour bed localisation following breast conserving surgery – Audit on behalf of the IMPORT trial management group. Eur J Surg Oncol (EJSO). 2009;35(6):578–82.
10. Hodapp N. Der ICRU-Report 83: verordnung, dokumentation und kommunikation der fluenzmodulierten photonenstrahlentherapie (IMRT). Strahlenther Onkol. 2012;188(1):97–100.
11. Moser EC, Vrieling C. Accelerated partial breast irradiation: the need for well-defined patient selection criteria, improved volume definitions, close follow-up and discussion of salvage treatment. Breast. 2012;21(6):707–15.
12. Dzhugashvili M, Pichenot C, Dunant A, Balleyguier C, Delaloge S, Mathieu M-C, Garbay J-R, Marsiglia H, Bourgier C. Surgical clips assist in the visualization of the lumpectomy cavity in Three-Dimensional conformal accelerated Partial-Breast irradiation. Int J Radiation Oncology*Biology*Physics. 2010;76(5):1320–4.
13. Ippolito E, Trodella L, Silipigni S, D'Angelillo RM, Di Donato A, Fiore M, Grasso A, Angelini E, Ramella S, Altomare V. Estimating the value of surgical clips for target volume delineation in external beam partial breast radiotherapy. Clin Oncol. 2014;26(11):677–83.
14. Coles CE, Brunt AM, Wheatley D, Mukesh MB, Yarnold JR. Breast radiotherapy: less is more?? Clin Oncol. 2013;25(2):127–34.
15. Kazemimoghadam M, Chi W, Rahimi A, Kim N, Alluri P, Nwachukwu C, Lu W, Gu X. Saliency-guided deep learning network for automatic tumor bed volume delineation in post-operative breast irradiation. Phys Med Biol. 2021;66(17):175019.
16. Buehler A, Ng S-K, Lyatskaya Y, Stsepankou D, Hesser J, Zygmanski P. Evaluation of clip localization for different kilovoltage imaging modalities as applied to partial breast irradiation setup. Med Phys. 2009;36(3):821–34.
17. Ng SK, Lyatskaya Y, Stsepankou D, Hesser J, Bellon JR, Wong JS, Zygmanski P. Automation of clip localization in digital tomosynthesis for setup of breast cancer patients. Physica Med. 2013;29(1):75–82.
18. Huang C, Han H, Yao Q, Zhu S, Zhou SK. 3D U2-Net: A 3D universal U-Net for Multi-Domain medical image segmentation. arXiv:1909.06012. 2019.
19. Feulner J, Kevin Zhou S, Hammon M, Hornegger J, Comaniciu D. Lymph node detection and segmentation in chest CT data using discriminative learning and a Spatial prior. Med Image Anal. 2013;17(2):254–70.
20. Xu J, Li M, Zhu Z. Automatic Data Augmentation for 3D Medical Image Segmentation. In: 2020 2020: Springer International Publishing; 2020: 378–387.
21. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B et al. Attention U-Net: learning where to look for the pancreas. arXiv:1804.03999. 2018.
22. Chen C, Biffi C, Tarroni G, Petersen S, Bai W, Rueckert D. Learning shape priors for robust cardiac MR segmentation from Multi-view images. arXiv:1907.09983. 2019.
23. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: 2016 2016: Springer International Publishing; 2016: 424–432.
24. Milletari F, Navab N, Ahmadi S-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV): 2016 2016; 2016: 565–571.
25. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:150504597 [cs] 2015.
26. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, Wasserthal J, Koehler G, Norajitra T, Wirkert S, Maier-Hein KH. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. arXiv:180910486 [cs] 2018.
27. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is All you Need.11.

28. Islam S, Elmekki H, Elsebai A, Bentahar J, Drawel N, Rjoub G, Pedrycz W. A comprehensive survey on applications of Transformers for deep learning tasks. arXiv:2306.07303. 2023.
29. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929. 2021.
30. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth H, Xu D. UNETR: Transformers for 3D medical image segmentation. arXiv:2103.10504. 2021.
31. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth H, Xu D. Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MRI images. arXiv:2201.01266. 2022.
32. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306. 2021.
33. Gao Y, Huang R, Yang Y, Zhang J, Shao K, Tao C, Chen Y, Metaxas DN, Li H, Chen M. FocusNetv2: imbalanced large and small organ segmentation with adversarial shape constraint for head and neck CT images. Med Image Anal. 2021;67:101831.
34. Tao R, Liu W, Zheng G. Spine-transformers: vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3D Transformers. Med Image Anal. 2022;75:102258.
35. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 2016-06 2016; 2016: 770–778.
36. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation networks. arXiv:1709.01507. 2019.

Publisher's note