Lin Ma, Junjie Wang, Shu Gong*, Libin Lan*, Li Geng, Siping Wang and Xin Feng

Self-supervised context-aware correlation filter for robust landmark tracking in liver ultrasound sequences

https://doi.org/10.1515/bmt-2022-0489
Received December 15, 2022; accepted January 5, 2024;
published online February 7, 2024

Abstract

Objectives: Respiratory motion-induced displacement of internal organs poses a significant challenge in image-guided radiation therapy, particularly affecting liver landmark tracking accuracy.
Methods: Addressing this concern, we propose a self-supervised method for robust landmark tracking in long liver ultrasound sequences. Our approach leverages a Siamese-based context-aware correlation filter network, trained by using the consistency loss between forward tracking and back verification. By effectively utilizing both labeled and unlabeled liver ultrasound images, our model, ***Siam-CCF***, mitigates the impact of speckle noise and artifacts on ultrasonic image tracking by a context-aware correlation filter. Additionally, a fusion strategy for template patch feature helps the tracker to obtain rich appearance information around the point-landmark.
Results: ***Siam-CCF*** achieves a mean tracking error of 0.79 ± 0.83 mm at a frame rate of 118.6 fps, exhibiting a superior speed-accuracy trade-off on the public MICCAI 2015 Challenge on Liver Ultrasound Tracking (CLUST2015) 2D dataset. This performance won the 5th place on the CLUST2015 2D point-landmark tracking task.
Conclusions: Extensive experiments validate the effectiveness of our proposed approach, establishing it as one of the top-performing techniques on the CLUST2015 online leaderboard at the time of this submission.

***Corresponding authors: Shu Gong**, Department of Gastroenterology, Children's Hospital of Chongqing Medical University, Chongqing, China, E-mail: chloe.gong@hotmail.com; and **Libin Lan**, College of Computer Science and Engineering, Chongqing University of Technology, Chongqing, China, E-mail: lanlbn@cqut.edu.cn. https://orcid.org/0000-0003-4754-813X (L. Lan)
Lin Ma, Junjie Wang, Siping Wang and Xin Feng, College of Computer Science and Engineering, Chongqing University of Technology, Chongqing, China, E-mail: malin@stu.cqut.edu.cn (L. Ma), junjie.wang@stu.cqut.edu.cn (J. Wang), Spwapex@stu.cqut.edu.cn (S. Wang), xfeng@cqut.edu.cn (X. Feng)
Li Geng, City University of New York NYCCT, New York, USA, E-mail: LGeng@citytech.cuny.edu

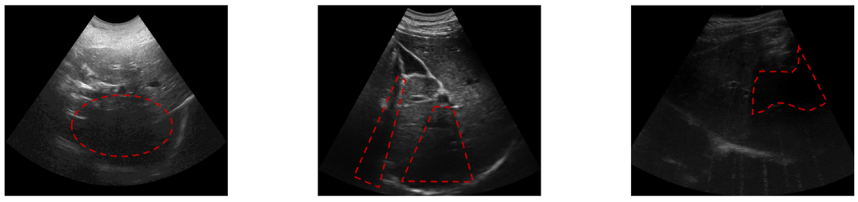
Keywords: self-supervised context-aware correlation filter; liver ultrasound landmark tracking; respiratory motion estimation; image-guided radiation therapy

Introduction

Image guide based precise conformal radiotherapy [1] has observably improved the effect of treatment of cancer patients. However, there are still some challenges associated with the technology. Among these challenges, the hardest thing to be coped with is the internal organ displacement [2] caused by respiratory motion, which will increase the uncertainties from breathing and drift during image-guided radiation therapy. That may negatively affect the accuracy and efficacy of the treatment [3]. Thus, it is necessary to compensate for respiration motion both accurately and at a real-time speed [4].

Ultrasound (US) imaging as an image-guided treatment protocol has shown significant advantages over other medical imaging methods such as computed tomography (CT) and magnetic resonance imaging (MRI) to guide treatment in radiation therapy. These advantages including better cost effectiveness, non-ionizing, and real-time imaging, make it one of the most ideal techniques for anatomical landmark tracking in liver ultrasound sequences. Further, accurate and robust motion tracking can benefit image-guided radiation therapy (IGRT) [5–7]. That ensures treatment quality, while reducing therapy margins and radiation exposures, and hence sparing healthy tissues. Thus recently, various methods using ultrasound images for respiratory motion tracking have been proposed, and achieved the most promising results [8–12]. However, it is still challenging to achieve high-performance landmark tracking using ultrasound images, owing to several major disadvantages, including speckle noise, artifacts, and blurred edge regions, as shown in Figure 1. For example, the speckle noise and artifacts make it extremely difficult to distinguish the characteristics of landmarks, which results in target drift during tracking (Figure 1(a) and (b)). Moreover, the blurred edge region leads to reduced appearance information of the area centered at the point-landmark, which degrades the performance of the tracker (Figure 1(c)).

The early tracking techniques based on liver ultrasound images mostly adopted block matching between adjacent



(a) (b) (c)

Figure 1: Significant challenges in liver ultrasound sequences. (a) Speckle noise. (b) Ultrasound artifacts. (c) Blurred edge region.

frames [13]. Some block matching approaches usually use normalized cross-correlation [12] or metric learning [14] as a similarity measure to localize the targets needed to be tracked in ultrasound sequences. Other approaches [10, 15] combine optical flow with block matching for motion estimation between frames. However, these methods typically rely on handcrafted features and extensive hyper-parameters, and thus unsuitable for complicated anatomical scenarios. Besides, due to the high speed of correlation filtering [16], it is also used to calculate the correlation between the target and subsequent frames in liver ultrasound sequences [17–19] as so to obtain the most relevant region in corresponding frame. But the quality of input features directly affects the final accuracy of model [20]. In recent years, using convolutional neural networks (CNNs) to liver ultrasound sequences tracking has achieved promising results [9, 11, 21]. However, there exit some problems when using CNNs to liver ultrasound tracking. Firstly, although pooling is a widely used operation for training deep model, pooling operation increases the uncertainty of ultrasound target tracking, which needs to predict pixel-level location. Secondly, in the field of medical imaging, directly getting enough available labeled medical images to train a network with high representation ability, is extremely difficult. While using insufficient data to train network may inhibit its learning ability, resulting in the underfitting of a complex model. So, attempting to evaluate the tracking performance of model based on a small number of labeled data, in a supervised way, is impractical and even infeasible.

All these problems motivate us to propose a self-supervised method for robust landmark tracking in long liver ultrasound sequences, as so to improve the quality of image-guided radiation therapy. Inspired by the recent success of using Siamese correlation filter network to visual tracking [22, 23], we intend to train a robust and effective tracker using consistency loss between the forward tracking and back verification, in a self-supervised manner. Besides, inspired by context-aware correlation filter [24], we build a Siamese-based context-aware correlation filter network, referred to as ***Siam*-CCF**, which is trained by fully taking advantage of both labeled liver ultrasound images and unlabeled ones. The context-aware correlation filter considers global information in the context, which mitigates the effects of speckle noise and artifacts in ultrasound images, and thus improves the robustness of model. Furthermore, we design a fusion strategy for template patch feature to help the tracker obtain rich appearance information of the area centered at the point-landmark. Meanwhile, we use CNN without pooling to extract the shallow features of the image, as so to retain more fine-grained feature information to improve the accuracy of target localization.

Quantitative and qualitative evaluations on the CLUST2015 2D dataset demonstrate that our proposed ***Siam*-CCF** yields a mean tracking error of [0.79 ± 0.83] mm at fps of 118.6, which proves that it can achieve a promising result and significantly outperform other methods with respect to real time.

The main contributions of this work can be summarized as follows:

1) We propose a self-supervised method for landmark tracking in liver ultrasound sequences using a Siamese structure network. His method makes full use of all data to train a tracking model with high generalization ability.
2) We introduce context-aware correlation filter into the proposed self-supervised framework. This reduces the effects of noise in ultrasound images, and thus improves the robustness to target appearance changes by explicitly incorporating context during the learning process.
3) We design a simple and efficient fusion strategy for template patch feature to selectively perform feature fusion, which is based on the current tracking results and the quality evaluation of response map.
4) We have conducted extensive comparative studies on the CLUST2015 2D dataset. Experimental results demonstrate that our method is effective and can achieve a promising result, particularly in real time, which has a prominent advantage in contrast to other methods.

The remainder of this paper is organized as follows. Section 2 reviews previous related work. Section 3 describes our approach and training process in detail. Section 4 gives experimental setup details, Section 5 gives ablation study, and comparison results on the CLUST2015 2D dataset, and finally Section 6 concludes the paper, and briefly discusses the limitations of our self-supervised learning method.

Related work

Ultrasound (US) imaging is a widely used medical imaging technique. With the rapid development of this technique, a large number of tracking methods based on ultrasound images, have been proposed. A common practice is to do similarity matching between cropped blocks of an image sequence [10, 12, 14, 15, 25]. Hallack et al. [25] use differentially isomorphic logDemons for region-of-interest image registration and dense Scale Invariant Feature Transform (SIFT) [26] as a similarity measure. Nouri and Rothberg [14] propose to map pixel intensity value of image block to low-dimensional space through a function, which uses the Euclidean distance metric, and find the position of the minimum distance metric from the template through window search. Shepard et al. [12] present an NCC-based block matching algorithm that simultaneously combines multiple templates to determine the affine transformation from previous frame to current frame, and applies many strategies to improve the accuracy and robustness of tracking in their framework. However, these methods rely on fine-tuning of a large number of hyperparameters, as so to make them suitable for an application-specific scenario.

Correlation filtering is based on fast Fourier transform (FFT), which can convert similarity calculation process in spatial domain into frequency domain, and exhibits high-speed performance in target tracking tasks. Literatures [17, 19] propose to apply Kernel Correlation Filter (KCF) to ultrasound image tracking, in which KCF is used to compute the similarity of plaques between both the current and subsequent frame in ultrasound sequences. And the position with the largest similarity is considered as new target tracking position. Kondo et al. [17] extend the KCF by using an adaptative window size and motion vector refinement with template matching to improve the tracking performance. Shen et al. [18] propose a robust tracker to minimize tracking error, which involves a scale adaptive KCF, an improved update rule, elaborately devising displacement and appearance constrains, and calculating a weighted displacement. Di et al. [27] present a Thermal Infrared (TIR) target tracking method, ASTMT, based on the Aligned Spatio-Temporal Memory Network. This approach models the TIR target tracking scene using spatio-temporal memory networks, reducing similarity interference. Simultaneously, an alignment matching module is utilized to enhance the model's robustness and tracking accuracy. Di et al. [28] introduce a self-supervised tracker, self-SDCT, within the deep correlation framework, employing multi-cycle consistency loss and similarity dropout strategies to achieve high-quality feature extraction and robust localization of tracked targets. These above-mentioned methods are competitive in real time, but their tracking performance is highly dependent on the quality of target template, and most of them use handcrafted features and even pixel intensities for tracking position prediction, which could not cope with the appearance changes of complex and variable plaque in ultrasound sequences.

It is well known that CNN technique has been successfully applied to a variety of visual tracking tasks. So, recently much work using CNN technique has also been devoted to the landmark tracking in ultrasonic sequence [9, 11, 21, 29]. The work proposed by Gomariz et al. [9] is the first to apply CNN to the liver ultrasound sequence tracking. It adopts a fully-convolutional Siamese network to track target features, and combines a location prior with a network-predicted location probability map to iteratively track targets in ultrasound images. Bharadwaj et al. [21] use an updated Siamese-based network approach for robust and accurate landmark tracking by introducing template update and a linear Kalman filter (LKF) into original architecture. To improve the tracking performance of CNN-based methods, Liu et al. [11] propose a cascaded Siamese network structure to improve the accuracy of landmark tracking through a strategy of coarse positioning to fine positioning. Similarly, Di et al. [30] propose an Adaptive Spatio-Temporal Context-Aware (ASTCA) model within the DCF-based tracking framework to enhance the accuracy of unmanned aerial vehicle (UAV) tracking and mitigate the impact of boundary effects. This model can learn spatio-temporal context weights, accurately distinguish targets from the background, and incorporate spatial context information in scenarios involving small targets and aerial views, effectively reducing background interference. Wu et al. [29] propose a fusion Siamese network with drift correction, in which four response maps generated by the cross-correlation were fused to reduce up-sampling error, and a correction strategy was used to revise target drift predicted by the network. However, training this CNN network needs enough labeled ultrasound images, while as we all know that is impractical for liver ultrasound landmark tracking task. Although Liu et al. [11] design an unsupervised training strategy, but it is too complicated and heavily depends on the quality of the selected corner points. Di et al. [31] introduces an active learning approach for deep visual tracking, aiming to train deep convolutional neural network models by selecting and annotating unlabeled samples. However, the challenge lies in the difficulty of choosing distinctive samples and dealing with the high computational complexity of the model.

Furthermore, some deep learning methods, have made much success in other visual tracking tasks [16, 22, 24, 32–37]. And researchers have explored the application of transformer technology in tracking tasks. Xin et al. [38] introduced

a Transformer tracking method (TransT) utilizing Siamese-like feature extraction backbone, attention-based fusion mechanism, and classification/regression head. Bin et al. [39] proposed a tracking architecture incorporating an encoder-decoder transformer; the encoder models global spatio-temporal feature dependencies, while the decoder predicts target object positions through query embedding. Botao et al. [40] introduced a novel one-stream tracking framework (OSTrack) unifying feature learning and relation modeling via bidirectional information flows in template-search image pairs. However, to the best of our knowledge, these existing approaches show that they almost all require a large number of ground truth labels for training, which is impractical. Thus, some self-supervised work using various pretext tasks as supervision is proposed [22, 32, 33, 37]. These pretext tasks involve colorization [33], cycle-consistency in time [32], or pseudo-labels [22]. **Our method closely relates to** the Siamese correlation filter network (CFN) [19, 22], which utilizes forward tracking and backward verification to train a self-supervised tracker without heavyweight annotations on public dataset of natural scene. Thus, we intend to use this idea to liver ultrasound sequences tracking tasks, and empirically show that our self-supervised learning approach is effective.

Methods

Our self-supervised framework for liver ultrasonic tracking is shown in Figure 2. We randomly select two frames, F_1 and F_2, from the video sequence. Firstly, in the previous frame F_1, we randomly initialize a target landmark and obtain a template patch z, centered around that landmark, with a size of 125×125. Then, using forward tracking, we predict the position of the target landmark in the subsequent frame F_2 and obtain a search patch x with a size of 125×125 centered around the predicted landmark. Finally, we reverse the tracking direction. We use the landmark position on the search patch x, predicted in the subsequent frame F_2, as a pseudo-label for backward validation. We then predict the position of the pseudo-label in the previous frame F_1. We expect the result of the backward tracking to match the initial landmark in the previous frame F_1, and we measure the difference between the forward and backward trajectories using a consistency loss trained by the network. We start by describing our self-supervised Siamese-based network framework and explaining how context-aware correlation filters [24] is applied to our self-supervised learning framework. We then elaborate on our proposed fusion strategy for template patch feature.

Siamese-based context-aware correlation filter network

In liver ultrasound sequence tracking scenarios, accurately and robustly tracking landmarks usually suffer from speckle noise, artifacts, and blurred edge regions. Directly applying discriminative correlation filter (DCF) [19, 23] to this tracking task could not achieve the promising results. This is because the DCF is unable to learn enough distinguishing information between both the targets and background, resulting in losing the tracked target in the test phase. To solve this problem, we incorporate context-aware correlation filter (CCF) [24] into the liver landmark tracking. The CCF explicitly learns a filter that has a high response to the template patch z in the current frame F_1 and close to zero response for context patches z_i. This is performed by adding the context patches z_i as a regularizer to the standard DCF formulation [22, 23] (see Eq. (1)). The context patches, denoted as z_i where $i \in [0, 1, \ldots, k]$, are patches of the same size as the template patch z. They are located

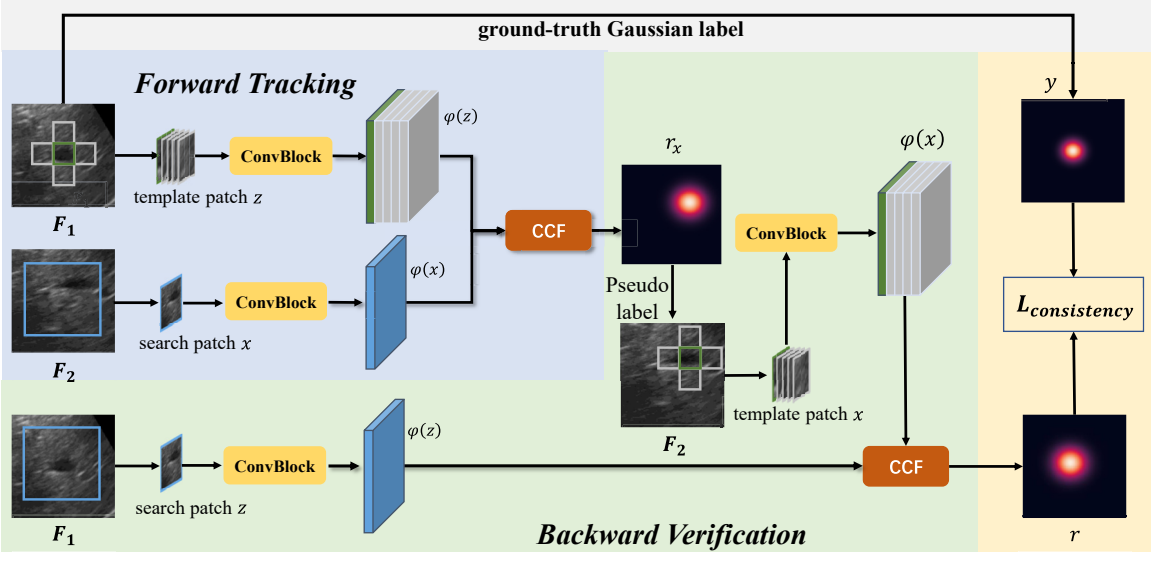


Figure 2: An overview of self-supervised ***Siam*-CCF** framework. In training phase, we train the tracker by forward tracking and backward validation. In inference phase, we used only the forward tracking. The blue background in the figure represents the forward tracking part of the model, and the green background represents the backward validation part of the model.

adjacent to the template patch z in the vicinity. And hence, it is beneficial to providing a closed-form solution to remain computationally efficient. This ensures our method has a good performance in real time. By minimizing Eq. (1), the desired filter w can be obtained.

$$\min_{w} \|w \star \varphi(z;\theta) - y\|^2 + \lambda_1 \|w\|^2 + \lambda_2 \sum_{i=1}^{k} \|w \star \varphi(z_i;\theta)\|^2, \tag{1}$$

where $\varphi(\cdot;\theta)$ denotes feature space encoded by parameters θ, $\star$ means circular correlation, λ_1 and λ_2 are constant regularization, y is an ideal Gaussian response map peaked at the target landmark, and k is the number of sampling patches. The following closed-form solution is obtained in the Fourier domain as follows [24]:

$$w = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\varphi(z;\theta)) \odot \mathcal{F}^{*}(y)}{\mathcal{F}^{*}(\varphi(z;\theta)) \odot \mathcal{F}(\varphi(z;\theta)) + \lambda_1 + \lambda_2 \sum_{i=1}^{k} \mathcal{F}^{*}(\varphi(z_i;\theta)) \odot \mathcal{F}(\varphi(z_i;\theta))}\right), \tag{2}$$

where $\odot$ is the element-wise product, $\mathcal{F}(\cdot)$ denotes the Discrete Fourier Transform (DFT), $\mathcal{F}^{-1}(\cdot)$ indicates the inverse DFT, and * is the complex conjugate operation.

We integrate the idea of CCF into the self-supervised Siamese-based network framework. **In the forward tracking**, the response map r_2 of a search patch x in the next frame F_2 can be obtained by convolving with the learned filter w:

$$r_x = \mathcal{F}^{-1}\left((\mathcal{F}(w_z))^{*} \odot \mathcal{F}(\varphi(x;\theta))\right). \tag{3}$$

In the backward verification, using the response map r_2, we create a pseudo Gaussian label denotes as y_2, and then generate a target template w_z by following Eq. (2). Finally, the response map r_1 corresponding to target patch x in the current frame F_1 can be computed by:

$$r = \mathcal{F}^{-1}\left((\mathcal{F}(w_x))^{*} \odot \mathcal{F}(\varphi(z;\theta))\right). \tag{4}$$

Ideally, the response map r should be identical with the originally given label y. In other word, the peak of r should be closed to the highest peak in the Gaussian response map y at the initialized landmark position. Thus, the network $\varphi(\theta)$ could be trained in a self-supervised manner by minimizing the following loss function as follows:

$$\mathcal{L}_{\text{consistency}} = \|r - y\|_2^2. \tag{5}$$

In the framework, we use a randomly initialized template patch in the current frame as label to predict its location in the next frame, in a forward track manner. Then, we reverse the sequence, and take the predicted location as a pseudo label to track backward. We formulate the difference between both initial template patch and the predicted patch as consistency loss for network training without ground truth annotations.

Fusion strategy for template patch feature

During correlation filter tracking, the template patch of the target changes continuously as tracking to adapt changes of the appearance in the target. In addition, the template patch is contaminated due to the accumulation of tracking offsets, which causes the tracker drifting. We empirically found that an appropriate template patch for the liver ultrasound sequence could improve the performance of correlation filter tracking. Therefore, we propose a fusion strategy for template patch feature to enhance the robustness of the ***Siam*-CCF**. Specifically, we obtain the current template patch based on the tracking results of the previous frame, which can adapt to complex changes in the appearance of the target, and since the initial template patch comes from the first frame of the ultrasound sequence, it correctly reflects the original appearance of the target. Then, we encode the initial and current template patches in the feature space to fuse them as a new feature.

To perform template feature fusion efficiently, we introduce the average peak-to correlation energy (APCE) [41] as a tracking confidence metric, which is defined as follows:

$$\text{APCE} = \frac{|\max(r) - \min(r)|^2}{\text{Mean}\left((r - \max(r))^2\right)}. \tag{6}$$

The response map r of the model output reflects the quality of the training sample (i.e., the feature of the template patch) during tracking. The response graph represents the predicted location probability of the target landmark in the template patch on the search patch. It has a distinct peak, where the highest response is centered around the location of the target landmark, and gradually decreases in value as the distance from the target landmark increases. When the tracking results are correct, the response map output by the model with a distinct peak and is close to the ideal 2-D Gaussian distribution. It also represents that the training samples used by the model accurately reflect the appearance of the target. Therefore, we fuse the features when the current APCE value is lower than its moving average. Based on Eq. (6), the improvement of our fusion strategy for template patch feature is as follows:

$$z_{\text{fusion}} = \begin{cases} \varphi(z_0,\theta) + \varphi(z_t,\theta), & \text{if APCE}_t < \beta \dfrac{1}{t-1} \displaystyle\sum_{i=1}^{t-1} \text{APCE}_i; \\ \varphi(z_t,\theta), & \text{others.} \end{cases} \tag{7}$$

where z_{fusion} is the fused feature, and β is a weight factor, z_0 and z_t represent the initial template patch of the first frame and template patch of the current frame, respectively.

Training process

In the CLUST 2D dataset, statistically, there are about 90 % of ultrasound sequence frames without annotations. Hence, we use self-supervised learning methods to train datasets with partially annotated and unannotated, which can improve the generalization and robustness of networks.

We consider that by initializing the target landmark at the center of the entire image, the target landmark will not move out of the search area in the short term. Based on this analysis, we randomly select three different frames from an ultrasound sequence with 10 consecutive frames, and initialize the target landmarks to obtain three patches of size 125 × 125 containing the target landmarks. One of which is selected as the template patch z, the remaining two are set as the search patch x, here we refer to the search patch x as x_{t1} and x_{t2} respectively. Tracking forward only once can lead to inaccurate target localization over longer tracking durations. Therefore, during training, we perform multiple forward traces, which helps to alleviate the problem of inaccurate target localization that may occur over time:

$$\begin{aligned} r_{x_{t1}} &= \mathcal{F}^{-1}\left((\mathcal{F}(w_z))^{*} \odot \mathcal{F}(\varphi(x_{t1};\theta))\right), \\ r_{x_{t2}} &= \mathcal{F}^{-1}\left((\mathcal{F}(w_{x_{t1}}))^{*} \odot \mathcal{F}(\varphi(x_{t2};\theta))\right), \\ r &= \mathcal{F}^{-1}\left((\mathcal{F}(w_{x_{t2}}))^{*} \odot \mathcal{F}(\varphi(z;\theta))\right). \end{aligned} \tag{8}$$

To train a model with stable performance, we use 3-frame temporal span in the actual training process, which requires that each

forward tracking and backward validation can correctly predict the landmark location, otherwise tracking failure errors will accumulate. Algorithm 1 shows the training iterations.

Algorithm 1: A train iteration of the tracking algorithm

Input: Template patch z and two search patches x_1, x_2, 2-D Gaussian map y_t
Output: The response map r

1 Extracting features of the input images, denoted as zf, xf_1, and xf_2
2 **Forward tracking**: calculate the corresponding response r_{t+1} = model(zf, xf_1)
3 Get the maximum value on the response map r_{x_1} and generate a pseudo-label for x_1 with this position
4 **Forward tracking**: calculate the corresponding response r_{t+2} = model(xf_1, xf_2)
5 Get the maximum value on the response map r_{x_2} and generate a pseudo-label for x_2 with this position
6 **Backward verification**: calculate the corresponding response r_t = model(xf_2, zf)
7 Calculate the consistency loss, loss = MSE(r_t, y_t)
8 Gradient back propagation

Inference

During inference, we only need the results of forward tracking, not backward verification. The annotation of the landmark position of the first frame is given in the video sequence. With this position as the center, an area is cropped as a template patch. The search patch is determined according to the target center position and target size.

Then, the template patch and the search patch are fed into the convolution block of ***Siam*-CCF** to extract features. The resulting feature map is calculated using a feature-level correlation filter to obtain a response map. For tracking the appearance variations in ultrasound sequence constantly, we update the filter parameters online as follows:

$$\mathcal{F}(w_t) = \gamma \cdot \mathcal{F}(w_t) + (1-\gamma) \cdot \mathcal{F}(w_{t-1}), \tag{9}$$

where $\gamma \in [0, 1]$ is the linear interpolation factor.

Experimental setup

Datasets

Our 2D liver ultrasound sequences are provided by the MICCAI 2015 Challenge on Liver Ultrasound Tracking (CLUST) [42] database. The data was gathered from ultrasound liver sequences of healthy volunteers while they were free breathing. The training set consists of 24 sequences, where 10 % of the images are annotated. The test set consists of 39 sequences, where the first frame of each sequence is provided with annotations. The sequence is divided into the following five groups (*CIL*, *ETH*, *ICR*, *MED1* and *MED2*) according to different sampling locations and scanners. The spatial resolution of the images ranges from 0.27 × 0.27 mm to 0.77 × 0.77 mm. The dataset contains 63 2D liver US sequences, characterized by full frames ranging from 895 to 15,640. The dataset was manually annotated by three observers and reviewed by another observer. The last comment for each target is the average of all three manual comments. Each ultrasound sequence has one to five landmarks. Details of the sequences are shown in Table 1 below. The tracking results will be submitted to the organizers, who will quantitatively evaluate the tracking results and the ground truth and display them on a leaderboard.

Table 1: Statistics for each category sequence in the CLUST dataset. We count the number of training and test sets for each category sequence, the average number of image frames per category sequence, and the average number of labels per category sequence.

Sequence class	Training set size	Test set size	Average image frames	Average labels frames
CIL	3	6	1,172	138
ETH	16	30	5,497	551
ICR	12	13	3,646	397
MED1	16	27	2,542	255
MED2	6	9	2,715	272

Evaluation criteria

Given the ground truth annotations p_i and tracking results $\hat{p}_i$ for target i, the tracking error (TE) is calculated as

$$TE_i(t) = \left|\left|p_i((t)) - \hat{p}_i(t)\right|\right|_2, \tag{10}$$

where $i \in \{CIL, ETH, ICR, MED1, MED2\}$, $\hat{p}_i(t)$ is the result of the considered tracking method tracking target i in frame t, and $p_i(t)$ is the result of marking target i in frame t using manual labeling. Eq. (10) is to use the Euclidean distance to measure the deviation between the tracking result and the manual labeling. Tracking error is summarized by the mean, standard deviation (Std), and 95th percentile (95th) of Euclidean distance [42] over all frames.

Implementation details

In our experiments, we have two ways to generate template patch z: without initial annotations and with initial annotations. In practice, three patches are cropped from the ultrasound sequence frames during the training process. Patch cropping position is the center point of sequence frames for those without initial annotations; for those with initial annotations, template patch z is center cropped with the annotation point, and search patch x_i is referenced to the position of template patch z annotation point. In our work, both approaches are referred to as self-supervised learning.

For the design of the network structure of the tracker, we refer to the work of Wang et al. [22] using DCFNet [23] as our baseline. Our network of the convolutional layer is comprised of Stage 1 of VGG-16 [43], removing the pooling layer and changing the output from 64 channels to 32 channels. Furthermore, we employ a local response normalization (LRN) layer at the end of the convolutional layers. In preprocessing training data, we merely crop the central patch of each frame. The patch size was 1/6 of the whole image and further adjusted to 125 × 125 as the network's input.

In the pre-processing process, we cropped and resized all the training data centers to the image of the specified size, 125 × 125. Then,

the ideal heatmaps (i.e., labels) are generated based on the center position using a Gaussian distribution with a sigma of 4. We train the model with the stochastic gradient descent (SGD) [44] optimizer, where the momentum is 0.9 and the weights decay to 0.005. Set the learning rate to 0.01 and train with 30 epochs and a mini-batch size of 64. We train all the network parameters from scratch without pre-training.

In the inference stage, we set a fixed scale transformation $S=\{s^n | s=1.0275, k=-1, 0, 1\}$ to obtain multiple scale response maps, and the location of the maximized response value in the response maps is the predicted landmark location. Empirically, we assign the model linear interpolation factor γ and the weight factor β to 0.01 and 1.0, respectively.

Our network was achieved with Pytorch [45], with the network training and the experiments ran on an Nvidia GeForce RTX 1080Ti GPU.

Results and discussion

Results

We evaluated our proposed model in the CLUST 2D test set. We evaluated our model on each image sequence group in the test set specified by the organizers. The image sequence group category of the test set is the same as that of the training set, the Sequence column shows the five categories of the image sequence group and overall, and the num column indicates the number of tests for each image sequence group. The overall accuracy of the model was 0.79 ± 0.83 mm and the 95th is 2.33 mm. Especially for the ETH category in a set of 30 image sequences, the accuracy reached 0.61 ± 0.50 mm, with a 95th percentile of 1.62 mm, as shown in Table 2.

Quantitative analysis

Table 2 shows the tracking performance of our method on the test set for different groups. The tracking error of mean is less than 1 mm for most sequences, with the *ETH* group performing best on all metrics. As shown in Figure 3, the median tracking error performance for each group remains low, and even with some outliers, the maximum outlier does not exceed 3.60 mm. It is worth mentioning that the *ETH* group has the highest number of sequence frames in the CLUST dataset but the lowest average tracking error, which indicates that our method can perform sufficiently robustly in ultrasonic tracking tasks under long tracking time.

Table 2: Quantitative result of landmark tracking on CLUST. The down arrow (↓) indicates that the smaller the number, the better.

Sequence	Num	Mean, mm ↓	Std, mm ↓	95th, mm ↓
CIL	6	1.24	1.23	4.29
ETH	30	0.61	0.50	1.62
ICR	13	0.88	1.04	3.36
MED1	27	1.11	1.16	3.58
MED2	9	0.91	0.80	2.54
Overall	85	0.79	0.83	2.33

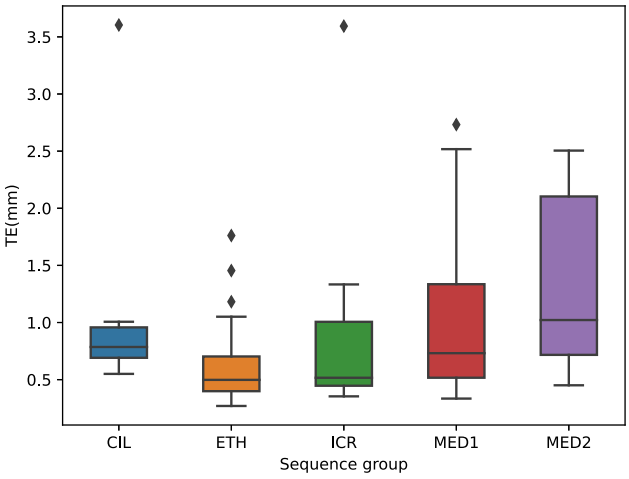


Figure 3: The tracking errors of the five groups on the test dataset obtained by ***Siam*-CCF**.

Table 3 shows our method compares to state-of-the-art methods in the test set. In Table 3, the no-tracking row indicates that no tracking method is used, and the method uses the landmark position on the initial frame to predict the landmark position on subsequent frames. This row points out the need for landmark tracking methods in image-guided radiotherapy. Furthermore, our method is much faster than the state-of-the-art methods in the inference phase, achieving a balance between accuracy and speed. Our method can also outperform the results of Shepard et al. and Williamson et al. in terms of the tracking error of Std, demonstrating the robustness of ***Siam*-CCF** on the liver ultrasound sequence landmark tracking task.

Table 4 shows the comparison of the tracking performance of our method with the state-of-the-art methods on different sets of test sets. In Table 4, except for the large performance difference in the *MED2* group, the performance differences of other groups are not significant. But in terms of processing speed, our model shows clear advantages.

Although the CLUST 2D dataset only provides a low proportion of scattered annotations, we can still make full use of it in our network training. Compared with self-supervised learning, we provide annotated landmarks in the first frame of training samples under self-supervised learning, which were set as template patches rather than random cropping patches. The search patches in the

Table 3: The performance of tracking test set for other state-of-the-art methods and our methods ([a] indicates the on-site challenge at MICCAI 2015 CLUST, no access to 20 % of all data before computation of the tracking result). The down arrow (↓) indicates that the smaller the number, the better.

Participant	Mean, mm ↓	Std, mm ↓	95th, mm ↓
Liu et al. [11]	0.69	0.67	1.57
Shepard et al. [12]	0.72	1.25	1.71
Williamson et al. [10]	0.74	1.03	1.85
Wang et al. [46]	0.75	0.62	1.65
Ours	0.79	0.83	2.33
Wu et al. [29]	0.80	1.16	2.29
Jeungyoon et al. [19]	0.85	0.80	2.32
Shen et al. [18]	1.11	0.91	2.68
Hallack et al.[a] [25]	1.21	3.17	2.82
Gomariz et al.[a] [9]	1.34	2.57	2.95
Makhinya and Goksel[a] [15]	1.44	2.80	3.62
Bharadwaj et al. [21]	1.60	3.69	4.21
Ihle [47]	2.48	5.09	15.13
Kondo [17]	2.91	10.52	5.18
Nouri and Rothberg [14]	3.35	5.21	14.19
No tracking [42]	6.25	5.11	16.48

Table 4: Comparison with state-of-the-art landmark tracking results in CLUT. The down arrow (↓) indicates that the smaller the number, the better.

Sequence	Method	Mean, mm ↓	Std, mm ↓	95th, mm ↓
CIL	Liu et al.	1.19	1.16	4.16
	Ours	1.24	1.23	4.29
ETH	Liu et al.	0.59	0.57	1.24
	Ours	0.61	0.50	1.62
ICR	Liu et al.	0.77	0.78	2.70
	Ours	0.88	1.04	3.36
MED1	Liu et al.	0.78	0.60	1.81
	Ours	1.11	1.16	3.58
MED2	Liu et al.	0.80	0.90	1.73
	Ours	0.91	0.80	2.54

Table 5: Performance of network trained with/without init annotations learning. The down arrow (↓) indicates that the smaller the number, the better.

Initial annotations	Mean, mm ↓	Std, mm ↓	95th, mm ↓
With	0.97	0.98	2.69
Without	1.19	1.14	2.92

training sample are determined based on the position of the patch template in the first frame. In this way, the patches of the input model can obtain more meaningful objects than randomly clipped ones. Table 5 shows the evaluation results, where self-supervised learning with initial annotations under the CCF resulted in a reduction of 18.49 , 14.04 and 7.88 % for TE, Std and 95th, respectively. The effect of this incompletely supervised training is most evident in the TE metric, which is also the most critical metric. Compared to randomly cropping patches, with initial annotations, which gives an initial target patch, allows the model to learn more meaningful tracking information and thus obtain better tracking results.

Qualitative analysis

To see the effect of our method more visually, we visualize some frames from the ultrasound sequence of the dataset. Figure 4 represents an example of tracking a landmark in a randomly selected sequence of images in the training set. For better illustration, we visualized the tracking results on some of the frames. See the bottom of Figure 4, the landmark locations predicted by our method are very close to the ground truth, and the tracking trajectory is similar to respiration showing a certain periodicity.

In order to emphasize the importance of the ***Siam*-CCF**, we visualize how it compares to the baseline in terms of accuracy on different image sequences. Table 6 has shown at the data level that CCF and template feature fusion strategies help improve the model's accuracy and robustness. Another aspect, in Figure 5, the tracking effects corresponding to the four image sequences are shown. Each row represents one image sequence. Column (a) represents the patches of the first frame of the image sequence. Column (b) represents the patches of the subsequent frames of the image sequence with ground truth. The first row represents an image sequence (CIL-01). Since the scenes of this sequence are not complex, the errors of both methods are within acceptable limits, but ***Siam*-CCF** is still better than the baseline. The second row represents an image sequence (ETH-01-1). The sequence frames in this group are generally extended, with the most prolonged frames being over 10,000. Although the long-time tracking task is difficult, our method is competent for this type of long-time tracking. The third row represents an image sequence (ICR-01), where the landmark template changes as it moves. Our fusion strategy for the template patch feature allows the model to obtain richer appearance information, keeping the template clean as the tracking progresses. The fourth row represents an image sequence (MED-02-3), in which only ***Siam*-CCF** can retrieve the landmark the baseline failed to do so. The baseline gradually shifts as the tracking

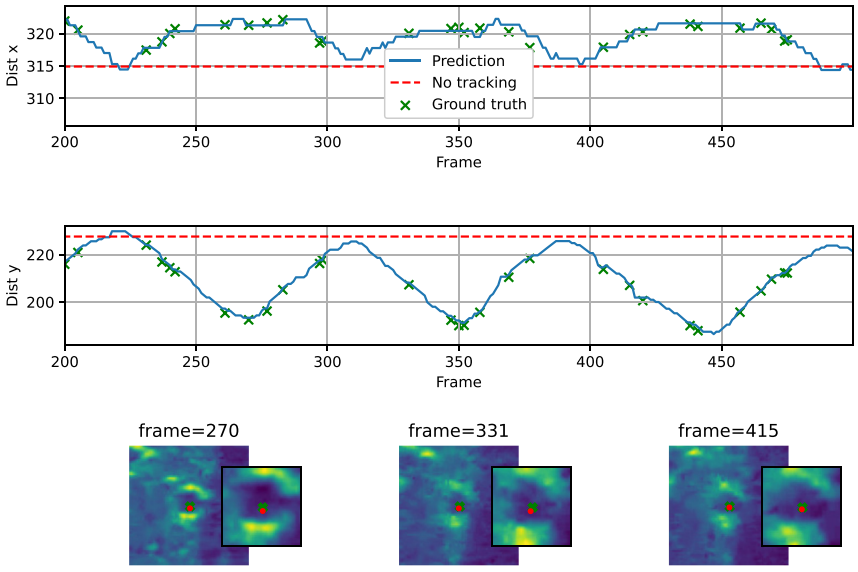


Figure 4: Results of tracking the part of the target ICR-04_1 (from 200 to 500 frames). (Top row) The corresponding result map of the predicted result (blue line) moving in the transverse direction of the image (Dim x) relative to the initial position (red dashed line) as the number of frames increases. The human-annotated ground truth $p_i(t)$ is shown as a green cross. (Middle row) The corresponding result map of the predicted result (blue line) moving in the longitudinal direction of the image (Dim y) relative to the initial position (red dashed line) as the number of frames increases. The human-annotated ground truth $p_i(t)$ is shown as a green cross. (Bottom row) Three images highlighting results in different frames, showing the differences between predicted $\hat{p}_i(t)$ (red dot) and ground truth $p_i(t)$ (green cross) positions.

Table 6: Ablation study with different modules in the network. Track error of mean, std, 95th and speed are reported. The down arrow (↓) indicates that the smaller the number, the better, and the up arrow (↑) indicates that the larger the number, the better.

Baseline	CCF	Fusion	Mean,- mm ↓	Std, mm ↓	95th, mm ↓	Speed, fps ↑
✓			1.49	1.57	4.26	241.4
✓		✓	1.33	1.29	3.78	202.1
✓	✓		1.31	1.27	3.24	138.7
✓	✓	✓	1.19	1.14	2.92	118.6

process progresses, and our context-aware filter keeps the tracking error within a low range.

Ablation study and analysis

In this section, we perform five cross-validation tests on the CLUST 2D dataset to verify the impact of different modules on the tracking performance. The configuration of the hyperparameters is the same unless otherwise specified.

Context-aware correlation filter

In our experiments, we explored the performance of context-aware correlation filters on the dataset. As shown in Table 6, experiments show that the CCF module can reduce the mean TE of landmarks, which significantly reduces the standard deviation (Std) and 95th percentile (95th) of TE. Specifically, without considering the fusion strategy for template patch feature, the mean TE is reduced by 0.18 mm with the CCF module, and the standard deviation (Std) and 95th percentile (95th) are reduced more significantly by 0.30 and 1.02 mm, respectively. As shown in Figure 6, the heatmap with the CCF module can focus more precisely on the landmark and suppress artifacts around the target from interfering with it. The above results indicate that the CCF module improves tracking robustness better than the baseline.

Fusion strategy for template patch feature

We experimentally verified the importance of the fusion strategy for template patch feature for our model, as shown in Table 6. Compared with non-template feature fusion methods, the tracking error with the fusion strategy for template patch feature is minor and more robust. The strategy mitigates the target drift during tracking by enriching the appearance information of the template, which reduces the overall tracking error and is less time-consuming.

The tracking error of our proposed method is significantly reduced compared to the baseline, as seen in the last row of Table 6, where the TE of mean, Std, and 95th metrics are reduced by 20.13 , 27.39, and 31.46 %, respectively. However, the fps of the inference stage decreases. After analysis,

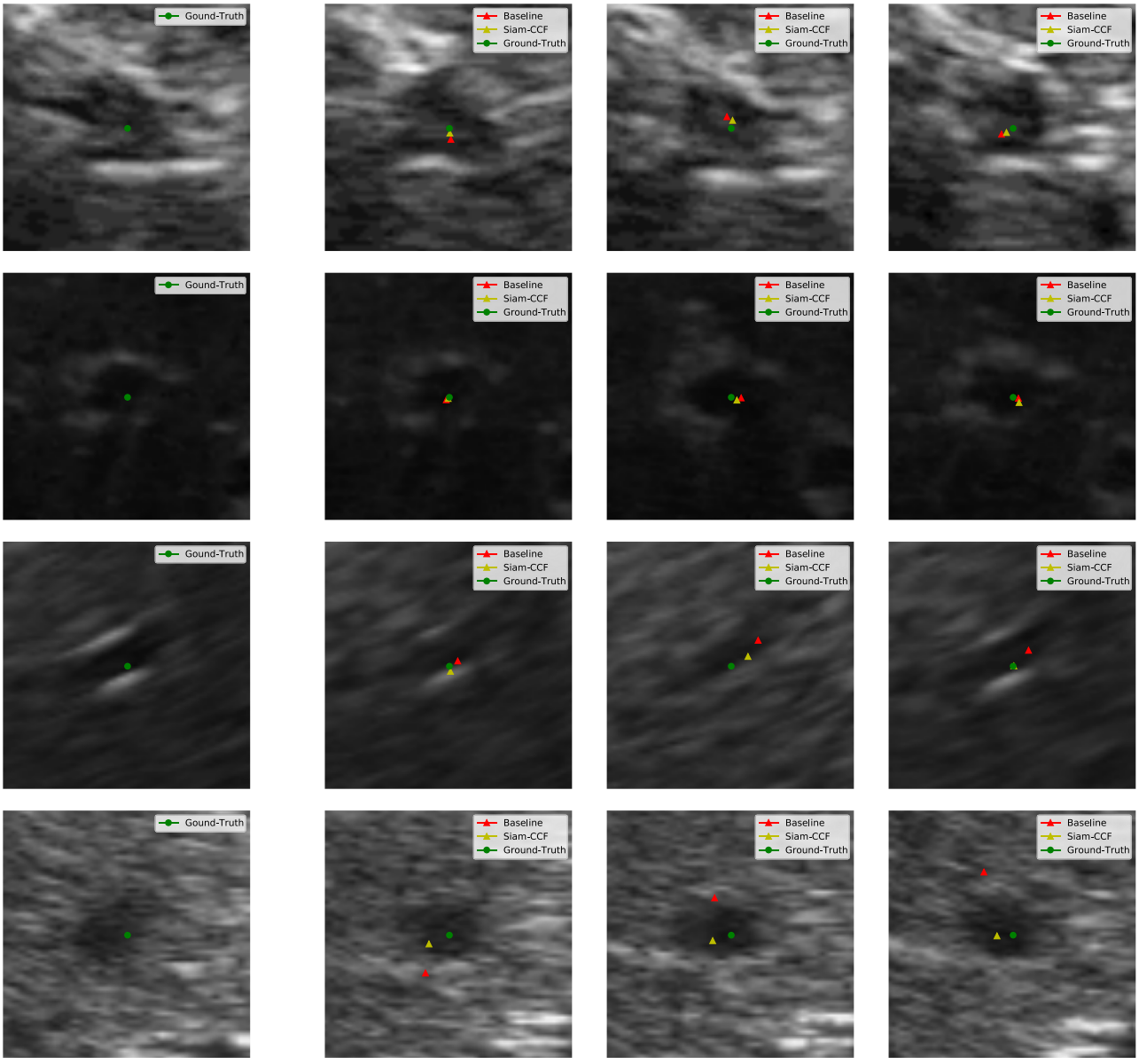


Figure 5: Visualization of baseline and ***Siam*-CCF** tracking results on four different sequence groups. Rows: patches of ultrasound image sequence. Columns: frames at different time periods and with annotations in a given ultrasound image sequence. First row: images from the first marker landmark of *CIL-01*. Second row: images from the first marker landmark of *ETH-01-1*. Third row: images from the first marker landmark of *ICR-01*. Fourth row: images from the fourth marker landmark of *MED-02-3*. (a) The first frame patch in the ultrasound image sequence, i.e. the template frame. (b) Search patch with ground-turth in subsequent image sequence

we find that CCF needs to calculate the filter responses around the template patches, and the fusion strategy for template patch feature increases the post-processing computation, which leads to the decrease of fps in the inference stage. Although adding modules will affect the inference speed of the model to some extent, our method can still maintain above 100 fps in inference speed due to the simplicity of our overall network structure.

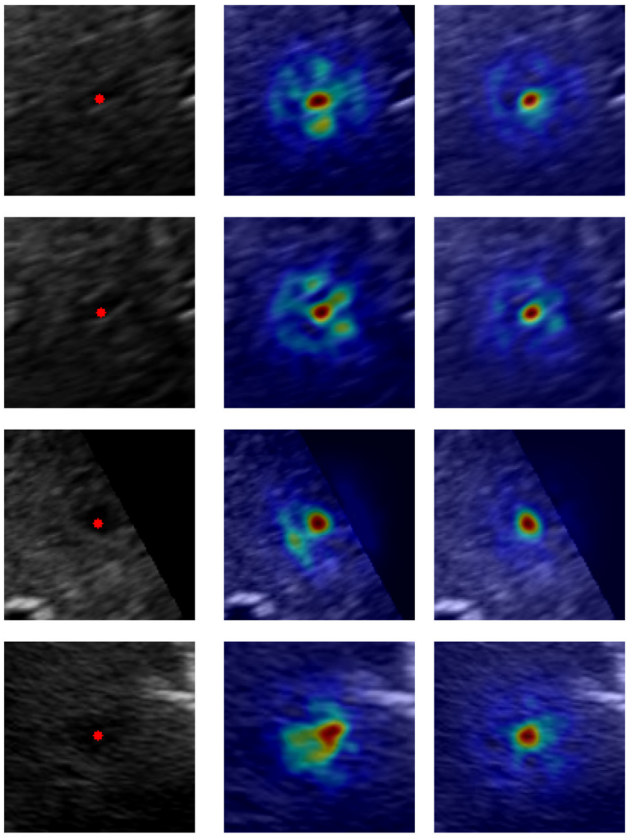


Figure 6: Visualization of the heatmap without/with CCF module tracking results. (a) Location of landmarks on search patch. (b) Tracking heatmap without CCF module. (c) Tracking heatmap with CCF module.

Conclusions

In this paper, we propose a robust, accurate, and efficient self-supervised tracking algorithm named as ***Siam*-CCF**. In ***Siam*-CCF**, we introduce a context-aware correlation filter to the siamese-based nerual network to implement self-supervised liver tracking. ***Siam*-CCF** is trained in a self-supervised manner using the loss of consistency between forward tracking and backward validation. Meanwhile, we propose a fusion strategy for template patch feature to help the model obtain richer template information. Our method achieves an overall accuracy of 0.79 ± 0.83 mm on the CLUST 2D dataset. Moreover, benefiting from the simplicity of the overall network structure, our tracker can easily achieve real-time tracking on the GPU. In comparison to models relying on annotated data, while our approach attains faster speeds and obviates the need for extensive annotation, this advantage is countered by a trade-off in accuracy. For our future endeavors, we plan to design a motion model integrating specific target motion patterns to enhance tracking accuracy.

Research ethics: The local Institutional Review Board deemed the study exempt from review.
Informed consent: Informed consent was obtained from all individuals included in this study.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.
Research funding: This work is supported in part by the Key project of the Chongqing Technology Innovation and Application Development under Grant No. cstc2021jscx-dxwtBX0018, and in part by the Natural Science Foundation of Chongqing under Grant No. CSTB2022NSCQ-MSX0493, and in part by the Chongqing Postgraduate Scientific Research Innovation Project under Grant No. CYS23678, and in part by the Action Plan for the High-quality Development of Postgraduate Education of Chongqing University of Technology under Grant No. gzlcx20233200, and in part by the Scientific Research Foundation of Chongqing University of Technology under Grant No. 0103210650, and in part by the Youth Project of Science and Technology Research Program of Chongqing Education Commission of China under Grant No. KJQN202301145.
Data availability: The raw data can be obtained on request from the corresponding author.

References

1. Mackie T, Kapatoes J, Ruchala K, Lu W, Wu C, Olivera G, et al. Image guidance for precise conformal radiotherapy. Int J Radiat Oncol Biol Phys 2003;56:89–105.
2. Langen K, Jones D. Organ motion and its management. Int J Radiat Oncol Biol Phys 2001;50:265–78.
3. Merwe D, Van Dyk J, Healy B, Zubizarreta E, Izewska J, Mijnheer B, et al. Accuracy requirements and uncertainties in radiotherapy: a report of the International Atomic Energy Agency. Acta Oncol 2017;56:1–6.
4. D'Souza WD, Naqvi S, Cedric X. Real-time intra-fraction-motion tracking using the treatment couch: a feasibility study. Phys Med Biol 2005;50:4021.
5. Brattain L, Telfer B, Dhyani M, Grajo J, Samir A. Machine learning for medical ultrasound: status, methods, and future opportunities. Abdom Radiol 2018;43:786–99.
6. Xing L, Thorndyke B, Schreibmann E, Yang Y, Li T, Kim G, et al. Overview of image-guided radiation therapy. Med Dosim 2006;31:91–112.
7. Wulff D, Kuhlemann I, Ernst F, Schweikard A, Ipsen S. Robust motion tracking of deformable targets in the liver using binary feature libraries in 4d ultrasound. Curr Direct Biomed Eng 2019;5:601–4.
8. Ha I, Wilms M, Handels H, Heinrich M. Model-based sparse-to-dense image registration for realtime respiratory motion estimation in image-guided interventions. IEEE Trans Biomed Eng 2018;66:302–10.
9. Gomariz A, Li W, Ozkan E, Tanner C, Goksel O. Siamese networks with location prior for landmark tracking in liver ultrasound sequences. In: 2019 IEEE 16th International Symposium On Biomedical Imaging (ISBI 2019); 2019:1757–60 pp.

10. Williamson T, Cheung W, Roberts S, Chauhan S. Ultrasound-based liver tracking utilizing a hybrid template/optical flow approach. Int J Comput Assist Radiol Surg 2018;13:1605–15.
11. Liu F, Liu D, Tian J, Xie X, Yang X, Wang K. Cascaded one-shot deformable convolutional neural networks: developing a deep learning model for respiratory motion estimation in ultrasound sequences. Med Image Anal 2020;65:101793.
12. Shepard A, Wang B, Foo T, Bednarz B. A block matching based approach with multiple simultaneous templates for the real-time 2D ultrasound tracking of liver vessels. Med Phys 2017;44:5889–900.
13. Giachetti A. Matching techniques to compute image motion. Image Vis Comput 2000;18:247–60.
14. Nouri D, Rothberg A. Liver ultrasound tracking using a learned distance metric. In: Proc. MICCAI workshop: challenge on liver ultrasound tracking; 2015:5–12 pp.
15. Makhinya M, Goksel O. Motion tracking in 2D ultrasound using vessel models and robust optic-flow. Proc MICCAI CLUST 2015;20:20–7.
16. Henriques J, Caseiro R, Martins P, Batista J. High-speed tracking with kernelized correlation filters. IEEE Trans Pattern Anal Mach Intell 2014; 37:583–96.
17. Kondo S. Liver ultrasound tracking using kernelized correlation filter with adaptive window size selection. In: Procedings MICCAI work. Chall. Liver ultrasound track; 2015:13–19 pp.
18. Shen C, Shi H, Sun T, Huang Y, Wu J. An online learning approach for robust motion tracking in liver ultrasound sequence. In: Chinese Conference On Pattern Recognition And Computer Vision (PRCV); 2018: 440–51 pp.
19. Shen C, He J, Huang Y, Wu J. Discriminative correlation filter network for robust landmark tracking in ultrasound guided intervention. In: International conference on medical image computing and computer-assisted intervention; 2019:646–54 pp.
20. Ma C, Huang J, Yang X, Yang M. Hierarchical convolutional features for visual tracking. In: Proceedings of the IEEE international conference on computer vision; 2015:3074–82 pp.
21. Bharadwaj S, Prasad S, Almekkawy M. An upgraded siamese neural network for motion tracking in ultrasound image sequences. IEEE Trans Ultrason Ferroelectrics Freq Control 2021;68:3515–27.
22. Wang N, Song Y, Ma C, Zhou W, Liu W, Li H. Unsupervised deep tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019:1308–17 pp.
23. Wang Q, Gao J, Xing J, Zhang M, Hu W. Dcfnet: discriminant correlation filters network for visual tracking; 2017. *ArXiv Preprint ArXiv:1704.04057*.
24. Mueller M, Smith N, Ghanem B. Context-aware correlation filter tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017:1396–404 pp.
25. Hallack A, Papiez B, Cifor A, Gooding M, Schnabel J. Robust liver ultrasound tracking using dense distinctive image features. In: MICCAI 2015 challenge on liver ultrasound tracking; 2015:28–35 pp.
26. Liu C, Yuen J, Torralba A. Sift flow: dense correspondence across scenes and its applications. IEEE Trans Pattern Anal Mach Intell 2010; 33:978–94.
27. Yuan D, Shu X, Qiao L, He Z. Aligned spatial-temporal memory network for thermal infrared target tracking. IEEE Trans Circ Syst II: Exp Briefs 2022;70:1224–8.
28. Yuan D, Chang X, Huang P-Y, Qiao L, He Z. Self-supervised deep correlation tracking. IEEE Trans Image Process 2020;30:976–85.
29. Wu C, Fu T, Wang Y, Lin Y, Wang Y, Ai D, et al. Fusion Siamese network with drift correction for target tracking in ultrasound sequences. Phys Med Biol 2022;67:045018.
30. Yuan D, Chang X, Li Z, He Z. Learning adaptive spatial-temporal context-aware correlation filters for uav tracking. ACM Trans Multimed Comput Commun Appl 2022;18:1–18.
31. Yuan D, Chang X, Qiao L, Yang Y, Wang D, Shu M, et al. Active learning for deep visual tracking. IEEE Transact Neural Networks Learn Syst 2023;1–13. https://doi.org/10.1109/tnnls.2023.3266837.
32. Wang X, Jabri A, Efros A. Learning correspondence from the cycle-consistency of time. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019:2566–76 pp.
33. Vondrick C, Shrivastava A, Fathi A, Guadarrama S, Murphy K. Tracking emerges by colorizing videos. In: Proceedings of the European Conference On Computer Vision (ECCV); 2018:391–408 pp.
34. Li P, Wang D, Wang L, Lu H. Deep visual tracking: review and experimental comparison. Pattern Recogn 2018;76:323–38.
35. LuNežič A, Vojíř T, Zajc L, Matas J, Kristan M. Discriminative correlation filter TracNer with channel and spatial reliability. Int J Comput Vis 2018; 126:671–88.
36. Muller M, Bibi A, Giancola S, Alsubaihi S, Ghanem B. Trackingnet: a large-scale dataset and benchmark for object tracking in the wild. In: Proceedings Of The European Conference On Computer Vision (ECCV); 2018:300–17 pp.
37. Meister S, Hur J, Roth S. Unflow: unsupervised learning of optical flow with a bidirectional census loss. In: Proceedings of the AAAI conference on artificial intelligence; 2018:32 p.
38. Chen X, Yan B, Zhu J, Wang D, Yang X, Lu H. Transformer tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021:8126–35 pp.
39. Yan B, Peng H, Fu J, Wang D, Lu H. Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021:10448–57 pp.
40. Ye B, Chang H, Ma B, Shan S, Chen X. Joint feature learning and relation modeling for tracking: a one-stream framework. In: European conference on computer vision. Springer; 2022:341–57 pp.
41. Wang M, Liu Y, Huang Z. Large margin object tracking with circulant feature maps. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017:4021–9 pp.
42. De Luca V, Banerjee J, Hallack A, Kondo S, Makhinya M, Nouri D, et al. Evaluation of 2D and 3D ultrasound tracking algorithms and impact on ultrasound-guided liver radiotherapy margins. Med Phys 2018;45: 4986–5003.
43. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition; 2014. *ArXiv Preprint ArXiv:1409.1556*.
44. Robbins H, Sutton M. A stochastic approximation method. Ann Math Stat 1951:400–7. https://doi.org/10.1214/aoms/1177729586.
45. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 2019;32:1–12.
46. Wang Y, Fu T, Wang Y, Xiao D, Lin Y, Fan J, et al. Multi3: multi-templates siamese network with multi-peaks detection and multi-features refinement for target tracking in ultrasound image sequences. Phys Med Biol 2022;67:195007.
47. Zachmann G, Frese I, Ihle F. Random forests for tracking on ultrasonic images [MS thesis]. Bremen, Germany: Univ. Bremen; 2017.