CacheTrack-YOLO: Real-time Detection and Tracking for Thyroid Nodules and Surrounding Tissues in Ultrasound Videos

Xiangqiong Wu, Guanghua Tan, Ningbo Zhu, Zhilun Chen, Yan Yang, Huaxuan Wen and Kenli Li

***Abstract*— To accurately detect and track the thyroid nodules in a video is a crucial step in the thyroid screening for identification of benign and malignant nodules in computer-aided diagnosis (CAD) system. Most existing methods just perform excellent on static frames selected by manual from ultrasound videos. However, manual acquisition is a labor-intensive work. To make the thyroid screening process in a more natural way with less labor operations, we develop a well-designed framework suitable for practical applications for thyroid nodule detection in ultrasound videos. Particularly, in order to make full use of the characteristics of thyroid videos, we propose a novel post-processing approach, called Cache-Track, which exploits the contextual relation among video frames to propagate the detection results into adjacent frames to refine the detection results. Additionally, our method can not only detect and count thyroid nodules, but also track and monitor surrounding tissues, which can greatly reduce the labor work and achieve computer-aided diagnosis. Experimental results exhibit our method performs better in balancing accuracy and speed.**

***Index Terms*— Deep learning, thyroid nodules, computer-aided diagnosis, ultrasound videos, detection and tracking.**

I. Introduction

THYROID gland, as an important endocrine tissue, can produce hormones that affect heart rate, body weight and control a multitude of other functions. Thyroid nodule is one of the most common thyroid gland diseases, which can cause serious damage to human health. The most popular imaging modality for scanning the thyroid is ultrasound imaging technique, because it possesses many advantages such as portability, low costs, real-time acquisition and harmlessness.

However, due to some limitations including the low image quality, image artifacts and speckle noise, especially the low contrast between the tissues and lesions, the neck tissues and lesions in ultrasound imaging are often blurred and unrecognizable. Moreover, the acquisition and the interpretation of ultrasound images heavily rely on the expertise and experience of the doctors, which is usually manual, subjective, time-consuming and tedious.

In order to improve diagnostic objectivity and reduce the workload of the radiologists, the methods of CAD have been developed for many years. However, there are still many challenges in automatic detection and recognition of tissues and lesions, such as inhomogeneity between thyroid gland and nodules, various sizes and shapes of lesions, etc. Nevertheless, artificial intelligence methods have made a lot of attempts to address various visual tasks [1] such as image classification [2], object detection [3], image segmentation [4] and action recognition [5]. In thyroid diagnosis, there existed many researches involving various features of ultrasound images to identify the locations and types of thyroid nodules.

Thyroid nodule classification in still images. Some early computer-aided thyroid diagnosis methods tended to feed hand-crafted feature sets into different classifiers. For example, Zhang *et al*. [6] adopted nine popular machine-learning algorithms to classify the types of thyroid nodules (benign or malignant) according to the eleven ultrasound hand-crafted characteristics such as echogenicity, margin, size and shape, etc. It proves that random forest algorithm outperforms the other machine learning algorithms and performs better than radiologists. But the machine learning algorithm needs to collect the features manually.

Since CNN models have indicated their robustness in extracting features from various images, Chi *et al*. [7] fine-tuned a CNN model to extract features from ultrasound images and put these features into a cost-sensitive random forest classifier to classify the types of thyroid ultrasound images. Meanwhile, Zhu *et al*. [8] exploited a new data augmentation method and the transfer learning to fine-tune a pretrained ResNet model that can classify the benign and malignant thyroid nodules based on ultrasound images. In addition, Liu *et al*. [9] combined deep semantic features extracted by a pretrained VGG model with conventional features to classify the thyroid nodules into benign and malignant nodules. Then, Song *et al*. [10] also utilized a pretrained CNN model to classify the types of thyroid nodules in the preprocessed ultrasound images, and the preprocess methods are limited to cropping the thyroid nodules in ultrasound images and resizing to a fixed

This work was supported by the National Natural Science Foundation of China (Grant No. 62072168) and National Key R&D Program of China (Grant No. 2019YFB2103005). *(Corresponding authors: Guanghua Tan, Kenli Li.)*

Xiangqiong Wu, Guanghua Tan, Ningbo Zhu, Zhilun Chen and Kenli Li are with the College of Computer Science and Electronic Engineering, Hunan University, China. (e-mail: joeyqwu@gmail.com, guanghuatan@hnu.edu.cn, quietwave@hnu.edu.cn, chenzhilun@hnu.edu.cn, lkl@hnu.edu.cn).

Yan Yang is with Department of Ultrasonic Diagnosis, The Second Affiliated Hospital and Yuying Children's Hospital, Wenzhou Medical University, Zhejiang, China. (e-mail: yang25yan@126.com).

Huaxuan Wen is with the Department of Ultrasound, Affiliated Shenzhen Maternal and Child Healthcare Hospital, Southern Medical University, Shenzhen, China. (e-mail: whxwell@126.com).

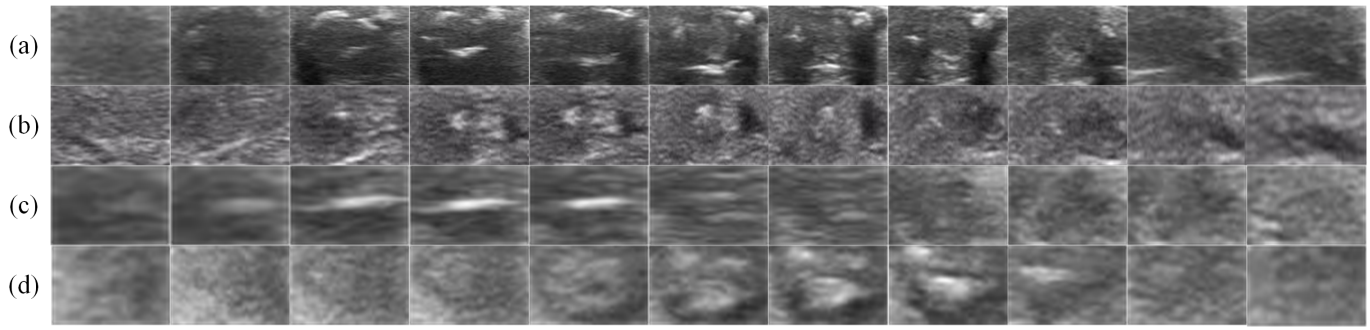

Fig. 1. Examples of partial frames of the thyroid nodules in ultrasound videos. Each row represents different frames of the same thyroid nodule. We cropped the thyroid nodules from ultrasound videos to demonstrate that the features of thyroid nodules are variable in ultrasound videos and it is difficult to locate thyroid nodules precisely except for the frames with obvious characteristics.

square size. Similarly, a multi-branch network with feature cropping was proposed in [11] to classify the thyroid nodules in ultrasound images. And Shi *et al*. [12] presented a novel knowledge-guided method to train a generative adversial network to augment the dataset to improve classification accuracy. Subsequently, they also introduced a multi-task cascade CNN model [13] to classify the thyroid nodules based on ultrasound images, which integrates the domain knowledge of the doctors and multi-modal ultrasound images to achieve more accurate thyroid nodule classification. Additionally, there existed a data-driven framework [14] based on belief rule base to conclude the doctors' diagnostic principles from historical examination reports, which can contribute to the diagnosis of new thyroid nodules. Although all these methods have very high accuracy in the classification of thyroid nodules, the first step in CAD system is to locate the positions of the thyroid nodules, which can reduce the workload of the doctors.

Thyroid nodule detection in still images. Object detection is one of the most popular tasks, which is the first step to implement an automatic CAD system. Most existing methods can be divided into two main types: anchor-free models [15]–[19] and anchor-based models [20]–[27]. Additionally, the anchor-based models contain two-stage models [20], [21] and one-stage models [22]–[27]. The anchor-free models often detect the objects based on keypoints and the anchor-based models need anchor boxes to filter the excessive potential bounding boxes. For two-stage models, the first stage is to propose a large number of regions of interest (ROIs) to classify the foreground and background by region proposal network (RPN), and the second stage is to refine the coordinates of bounding boxes and assign the class label to each bounding box. Therefore, the two-stage models can achieve better detection accuracy. However, it is also time-consuming to split two stages to detect the objects. One-stage models directly predict the coordinates of bounding boxes and class labels with a single network, which is more effective than two-stage models. However, although the one-stage models are faster than two-stage models, it is difficult for one-stage models to outperform the two-stage models in accuracy without RPN.

In order to locate the thyroid nodules in static images automatically and precisely, Ma *et al*. [28] applied a hybrid cascade CNN with two different CNNs to distinguish thyroid nodules from ultrasound images. Like two-stage models, it employed the first CNN to obtain the segmentation maps of thyroid nodules and the second CNN to detect thyroid nodules through splitting segmentation maps into different connected regions. Li *et al*. [29] also explored an improved two-stage detector to detect thyroid papillary carcinoma in ultrasound images, and thyroid papillary carcinoma is one of the most common malignant thyroid nodules. Specifically, to further boost the performance of smaller thyroid nodule detection, it combines a spatial constrained layer with a two-stage detector [21] to extract the surrounding region of thyroid cancer and make low-level image features and high-level semantic information concatenated, which improves the detection performance. Song *et al*. [30] developed a multi-task cascade CNN model to detect and recognize the thyroid nodules with a coarse-to-fine manner, which is also like a two-stage framework. The first stage is to detect the nodule locations based on a multi-scale SSD network, and the second stage is to refine the thyroid nodule classification. Liu *et al*. [31] introduced attention mechanism into Faster RCNN [21] and utilized the area around thyroid nodules as spatial prior information to improve the accracy of thyroid nodules localization. Although all methods mentioned above can achieve very high accuracy in detecting the thyroid nodules, they perform the detection and classification based on ultrasound static images. Moreover, the two-stage detectors still have a long way to satisfy the real-time requirement.

To accelerate the detection speed, Wang *et al*. demonstrated that an end-to-end still-image detection network [32] based on yolov2 [23] detected and classified thyroid nodules automatically and instantly. Although yolov2 [23] can run significantly fast, it still needs improvement before it can be directly applied to the video detection. If the temporal information and contextual information are not utilized well, the false negatives and false positives are prone to occur.

Even though all the methods mentioned above have achieved great success in the classification and detection of thyroid nodules, little attention has been devoted to detecting and tracking thyroid nodules and surrounding tissues based on thyroid ultrasound videos. Because the clinicians make a diagnosis based on the entire ultrasound scanning video instead of a single frame, and the acquisition of a single frame needs labor-intensive work, it is more imperative to assist doctors in real-time detection and monitoring based

on thyroid videos, which can reduce the labor intensity of the doctor and accelerate the diagnostic speed. In thyroid video detection, according to the initial documents retrieving, previous works only seek still-image detection, and we are the first to focus on video detection and tracking of thyroid nodules and surrounding tissues.

Thyroid nodule segmentation in still images. In addition, automatic segmentation of normal thyroid gland, thyroid nodules and cystic components [33] was applied to mobile health monitoring and volume estimation, which only needs a simple CNN with dilated convolutional layers to achieve an effective performance. However, this method also segments objects based on the ultrasound images rather than ultrasound videos. Moreover, an automated active contour model with intuitionistic fuzzy clustering was designed in [34] to segment thyroid nodules on ultrasound images. It eliminates human intervention and automatically identifies the thyroid nodules based on hypoechogenic characteristics. However, as shown in Fig. 1, since the echogenic features are not clear, it is difficult to accurately delineate the thyroid nodules in ultrasound videos.

Similarly, another segmentation method [35] utilized the echogenicity information in speckle-related pixels and local phase-based methods to segment multiple organs in ultrasound images, which can be a general framework for ultrasound-guided interventions and volumetric diagnosis. But they only segment the normal organs. For the speckle, an effective speckle reduction method [36] was proposed to improve the ultrasound image quality and enhance the visualization of tissue and lesion features, thereby facilitating segmentation and image analysis. Additionally, Feigin $et\ al.$ [37] presented a work toward single sided full waveform inversion in the ultrasound domain, which is the first deep learning framework to recover the sound speed maps from the plane wave ultrasound channel data. Although the results are encouraging, more research is needed to verify and improve this method.

Furthermore, a marker-guided U-net [38] combined with the doctors' mark information was proposed to solve the time-consuming process and difficult feature extraction of existing thyroid nodule detection methods. Although this method can achieve better performance, it still needs to manually mark four points to indicate the location of the thyroid nodule. Another segmentation approach [39] based on ultrasound image texture feature extraction was presented to show the value of a new parametrical model for texture characterization. This model can segment the thyroid by distinguishing thyroid and non-thyroid texture regions. Meanwhile, the authors [40] also combined this texture extraction approach with three different machine learning algorithms to classify and segment the thyroid gland. However, this approach based on the texture patch classification fails to locate precisely the edge between two tissues and always produces non-smooth boundary in ultrasound image.

Similar to ultrasound image detection, an epileptic seizure detection method was proposed in [41] to classify the electroencephalogram into several categories through an ensemble decision tree classifier, which extracts rich features from temporal, spectral and temporal-spectral domains to improve the classification accuracy. But it requires the data should be artifact-free. Similarly, Kordestani $et\ al.$ [42] conducted a systematic survey on failure diagnosis and prognosis. They classified the failure prognosis approaches into three groups: model-based, data-driven and knowledge-based methods. Although these various failure prognosis methods have achieved high accuracy, many challenges still need to be resolved in order to become more applicable to real-time systems.

Video Detection and Tracking methods. As to video detection and tracking in natural environments, most existing methods can also be categorized into two classes: tracking-by-detection methods and joint detection and tracking methods. The tracking-by-detection methods often seperate detection and tracking and utilize the tracking strategy as the post-processing approach of detection, when the joint detection and tracking methods learn jointly tracking association infomation and detection and implement the detection and tracking simultaneously.

In the tracking-by-detection paradigm, Seq-NMS [43] incorporated temporal information into the post-processing phase of still-image detection to find the best detection path according to high scores by using dynamic programming within the same clip. However, it is easy to add many false positives and make the detector drift from one object to another. Similarly, Sort [44] adopted a two-stage detection model for frame-to-frame prediction and utilized the kalman filter for motion prediction and the hungarian algorithm for data association. But Sort does not work well in the occlusion scene, and it is easy to switch the identity.

In order to obtain better performance in terms of identity switching, DeepSort [45] incorporated appearance information into data association metric based on Sort. In addition, Kang $et\ al.$ [46] proposed T-CNN that utilized the temporal and contextual information and optical-flow to suppress false positives and propagate results to adjacent frames, which is an extension of Faster-RCNN [21]. Then, Bergmann $et\ al.$ [47] also proposed Tracktor that replaced the region proposal network (RPN) with the tracking module, which can convert the detector into tracker without specific training and work online. Although these methods are accurate in detection, their backbone model Faster RCNN [21] is not suitable for video detection and tracking in real-time.

To further achieve the real-time performance, one-stage detection with tracking methods have been developed into the main stream. Wang $et\ al.$ proposed the JDE model [48] that incorporated the appearance embedding model into one-stage model Yolov3 [24] to obtain the detection results and corresponding embeddings simultaneously. Zhang $et\ al.$ and Zhou $et\ al.$ also presented real-time video tracking models FairMOT [49] and CenterTrack [50] based on anchor-free detection model, respectively. Although the joint detection and tracking methods are challenging and faster than the tracking-by-detection paradigm, the tracking-by-detection models are more stable than joint detection and tracking methods on the practical system. Our method belongs to the tracking-by-detection paradigm.

As demonstrated in Fig. 1, the features of thyroid nodules are always variable. Our datasets only contain detec-

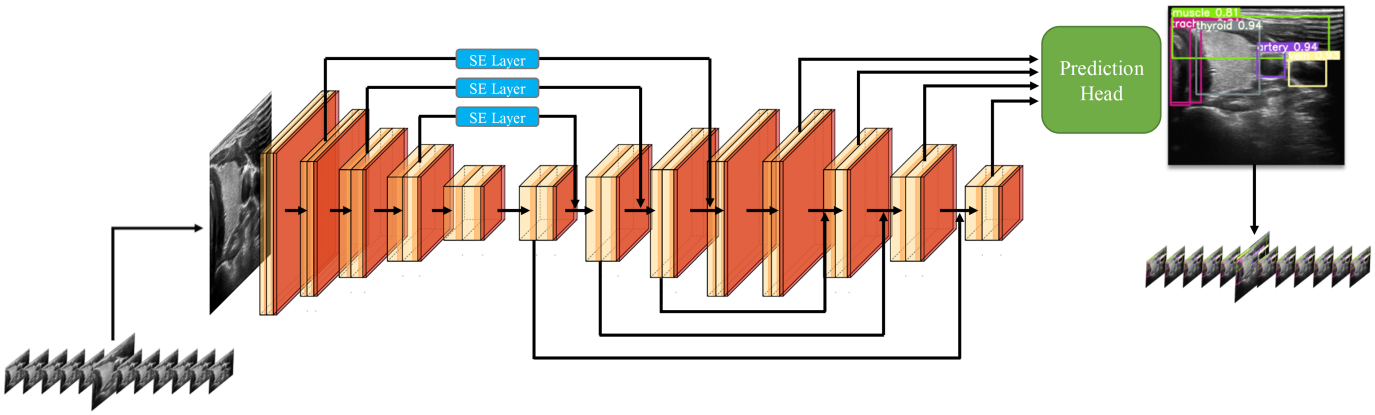

Fig. 2. The main architecture of proposed detection model. We adopt more low-level features into FPN and PAN to improve the accuacy of small-scale objects. We also introduce the squeeze-and-excitation layer (SELayer) that utilizes the channel-wise features to enhance the representation of network.

tion annotations and no tracking association annotations are available. Additionally, real time is a vital requirement to thyroid ultrasound video detection. As a consequence of these reasons, existing video detection and tracking models and still-image detections are not applicable to our thyroid ultrasound video dataset. Moreover, in thyroid diagnosis, in addition to thyroid nodules (TN), surrounding tissues including thyroid gland (TG), carotid artery (CA), cervical trachea (CT), internal jugular vein (IJV), cervical esophagus (CE) and anterior cervical muscles (ACM) are also crucial to diagnose the physical condition of the patient.

Therefore, inspired by the continuity of video frames and the characteristics of thyroid nodules in ultrasound videos, we focus on designing a real-time ultrasound video detector and tracker to detect and monitor multiple thyroid nodules and surrounding tissues in the transverse and longitudinal videos of thyroid gland, which can not only assist the doctors in making a diagnosis and reducing the workload from tremendous videos, but also make a preparation for classification of thyroid nodules in ultrasound videos.

In summary, our main contributions include the following parts:

1) We propose a real-time detection and tracking framework that can imitate the doctors' diagnostic thinking to reduce the doctors' workload and simplify the diagnosis process, which can jointly perform the detection and provide the statistics to the doctor in real-time manner. To our knowledge, our method is the first one in the thyroid literature to detect and track thyroid nodules and surrounding tissues in thyroid ultrasound videos.
2) We develop a multi-scale detection model based on the popular one-stage still-image detection framework (Yolo), which improves detection performance by integrating many effective components into the network, which can not only reduce the abundant features and the amount of computation, but also enhance the feature propagation and receptive fields of the model.
3) We design a novel cache track strategy that utilizes the contextual relation and temporal information among video frames to propagate the detection results into the adjacent frames and suppress the false positives among detection results, which can automatically modify the mistakes among detection results without sacrificing the speed of the detection network.

II. Methodology

In this section, we first illustrate the details of the improved detection network, which extends one-stage still-image detection network to perform robust by introducing useful components. Then we introduce the proposed cache track algorithm, which adopts the contextual information and temporal relation of video to modify the false detection results by simulating the doctors' diagnosis process.

A. Detection Network

The framework of the proposed detection network is presented in Fig. 2. Inspired by one-stage still-image detection framework [24], we employ darknet-53 [24] with cross stage partial (CSP) component [51] as the backbone network, which is similar to [25] and [26]. Because CSP uses partial transition layer and partial dense block as well as the split-merge strategy across stages, it can not only reduce the amount of computation caused by redundant features, but also improve the learning ability of network. [25], [26] and [51] have demonstrated the advantages and performance of CSP.

In the neck part of the proposed model, in addition to the feature pyramid network (FPN) [52] inherited from Yolov3 [24], we also adopt the spatial pyramid pooling (SPP) [53] and path aggregation network (PAN) [54] to improve multi-scale receptive fields and aggregate multi-layer feature maps, which is benefit to improve the robustness and recognition ability of the detection model. SPP was proposed in [53] to solve the problem of fixed-size images and generate multi-scale feature maps to improve the robustness of the model, because SPP can use multiple pool sizes to increase receptive fields, which can extract multi-scale contexts and improve detection accuracy. And PAN [54] is used to augment the path of information flow, which makes the feature propagation path from bottom level to top level shorten.

For object detection, low-level image features contain more detailed information, which is effective for improving identification about small-scale objects. Hence, to capture more low-level image features for small-scale objects, such as TN, CE and IJV, we include the feature maps of the second stage into the information flow path on the basis of the existing feature maps of [24]. Additionally, as shown in Fig. 2, we also employ the SELayer to selectively enhance the useful information. Specifically, SELayer proposed in [55] can generate normalized weighted map and assign weights to the channel-wise features by using adaptive average pooling and fully-connected layers.

The prediction head is the same as the one of Yolov3 [24]. It predicts the bounding box attributes, object confidences and classes simultaneously. The prediction head use the non-maximum-suppression (NMS) to remove the redundant bounding boxes. After NMS, the remaining bounding boxes will be put into the cache track module.

In our detection network, each image is labeled with a ground truth class label and a ground truth bounding box, and our detection result contain a predicted class label, a predicted target confidence and a predicted bounding box. In order to get better performance, we utilize the following loss function to train our model:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{conf} + \mathcal{L}_{box} \tag{1}$$

In class loss $\mathcal{L}_{cls}$ and confidence loss $\mathcal{L}_{conf}$, we adopt the binary cross entropy (BCE) loss to calculate the distance between the prediction and the ground truth. Specially, we associate the predicted target confidence with the Intersection-over-Union (IoU) calculated by the Eq. (5) to obtain the confidence loss. Therefore, each loss can be formulated as:

$$\mathcal{L}_{cls} = \sum_{i=1}^{n} \mathcal{H}(p_{cls}, B_{cls}^{gt}) \tag{2}$$

$$\mathcal{L}_{conf} = \sum_{i=1}^{n} \mathcal{H}(p_{conf}, \mathcal{L}_{CIoU}) \tag{3}$$

$$\mathcal{L}_{box} = \frac{1}{n} \sum_{i=1}^{n} (1 - \mathcal{L}_{CIoU}) \tag{4}$$

Where $\mathcal{H}$ means the binary cross entropy loss and B_{cls}^{gt} denotes the ground truth class, p_{cls} and p_{conf} represent the predicted class and confidence, respectively. In bounding box loss, we introduce [56], [57] to calculate the IoU between the predicted bounding box and the ground truth:

$$\mathcal{L}_{CIoU} = 1 - \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha * v \tag{5}$$

Here, $B = (x, y, w, h)$ represents the coordinate of the predicted bounding box, and $B^{gt} = (x^{gt}, y^{gt}, w^{gt}, h^{gt})$ denotes the coordinate of the ground truth, b and b^{gt} are the center point of B and B^{gt}, and ρ and c define the euclidean distance between the two center points and the diagonal length of the smallest enclosing box covering the two bounding boxes, respectively. In addition, $\alpha * v$ measures the consistency of aspect ratio, and α is a positive trade-off parameter,

$$\alpha = \frac{v}{(1 - \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|}) + v} \tag{6}$$

Where v denotes the consistency of aspect ratio,

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2 \tag{7}$$

Algorithm 1 Cache track algorithm

Input: Video sequence as ordered list $I = [i_0, i_1, ..., i_{T-1}]$ of images i_t, Detections as ordered list $\mathcal{D} = [D_0, D_1, ..., D_{T-1}]$ with $D_i = \{d_{t_1}^i, d_{t_2}^i, ..., d_{t_k}^i | 0 \leq t_1, ..., t_k \leq T-1\}$ as a list of ordered object bounding boxes $d_t^i = (x, y, w, h, classname, confidence)$, θ as IoU threshold.

Output: Set of matched trackers $\mathcal{T} = \{T_1, T_2, ..., T_m\}$ with $T_m = \{B_{t_1}^m, B_{t_2}^m, ..., B_{t_k}^m | 0 \leq t_1, ..., t_k \leq T-1\}$ as a list of ordered object bounding boxes $B_t^m = (x, y, w, h, trackerid, classname, confidence)$, Hit queue as ordered list Q=[$F_{t_1}, F_{t_2}, ..., F_{t_k} | 0 \leq t_1, ..., t_k \leq T-1$ and $F = 0/1$] of each tracker.

for $i = 0$ to $T-1$ **do**
  for i_t,$D_t \in$ zip($I, \mathcal{D}$) **do**
    Calculate the IoU distance d between detections D_t and trackers T_m;
    if $d > \theta$ and $D_t.classname$==$T_m.classname$ **then**
      Update matched trackers with assigned detections $T_m \Leftarrow D_t$ and $T_m.Q[-1] \leftarrow 1$;
    else
      Initialize a new tentative tracker $T_{tentative}$ for unmatched detection D_t and $T_{tentative}.Q[-1] \leftarrow 1$;
      Update the hit queue $T_m.Q[-1] \leftarrow 0$ for unmatched tracker T_m and change unmatched tracker into tentative tracker $T_{tentative} \Leftarrow T_m$;
    end if
    if $t \geq 7$ **then**
      for $T_m \in \mathcal{T}$ **do**
        if sum($T_m.Q$)>1 and sum($T_m.Q$)< len($T_m.Q$) and $T_m \in T_{active}$ **then**
          Complement the missing bounding boxes of frame t where $T_m.Q[t]$=0 using Kalman filtering prediction results;
          Update $T_m.Q[t] \leftarrow 1$;
        end if
      end for
    end if
  end for
end for

B. Cache Track

Accurate localization is a prerequisite for segmentation, tracking, and classification. Although the tracking-by-detection paradigm strongly depends on the performance of detection model, it is by far the most effective and stable method. In our method, the proposed cache track algorithm modifies the post-processing phase of detection model, which

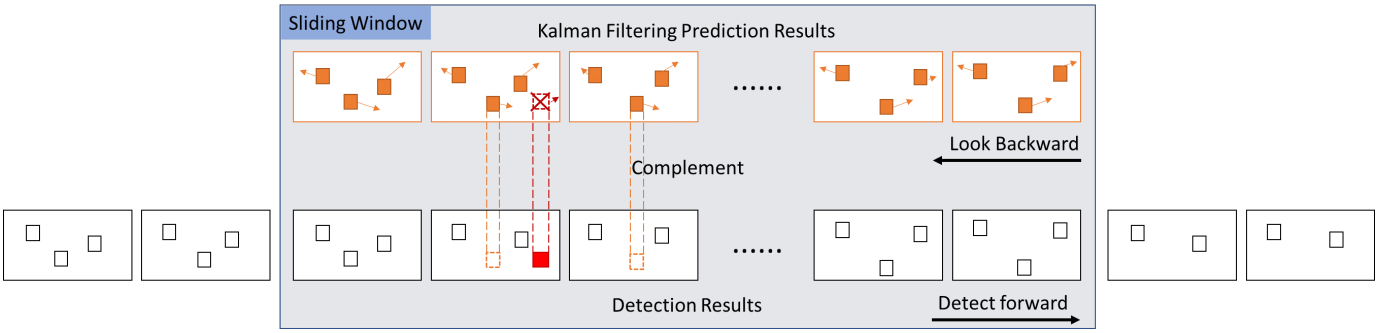

Fig. 3. The illustration of Sliding Window. It represents when the detector utilizes the sliding window to detect forward on the video sequence, the cache track strategy looks backward in the sliding window to complement the false negatives and suppress the false positives.

uses the kalman filter to predict next position on the basis of current position and hungarian algorithm to associate the current position with previous position.

However, when the still-image detection models are directly applied to video detection in a frame-by-frame manner, the false positives and false negatives are prone to generate. Even in adjacent frames, the detection results may vary greatly.

In order to alleviate this phenomenon, we use the eight frames as the width of the sliding window to obtain the time for subsequent frames to look backward in the sliding window. Unlike a majority of real-time tracking methods that can only detect forward, when our detection model utilizes the sliding window to detect forward on the video sequence, our cache track algorithm will imitate the doctors' diagnosis process to look backward in the sliding window to correct the false positives and false positives of the detection results. In this study, eight is just an empirical value.

As shown in Fig. 3, assuming that the bottom ordered sequence is the frame sequence in a video, and the small boxes in each rectangle represent the detection results of each frame. The gray rectangle represents the sliding window, and the orange boxes of the top sequence in the sliding window denote the kalman filtering prediction results. The false positive occurs in the fourth frame, when the detector detects the object as shown in the red box of fourth rectangle in the bottom sequence, the cache track will initialize a new tracker and kalman filtering will predict the position in the next frame and velocity for the red box. However, in the subsequent frames, the detector can not detect the object at the position predicted by kalman filtering, and the object will be regarded as the false positive and suppressed by the cache track algorithm. The false negatives occur in two consecutive frames, as illustrated in the orange dashed boxes of the fourth and the fifth frame of the bottom ordered sequence, and the detector fails to detect the objects in these two frames.

For this situation, the proposed cache track algorithm utilizes the kalman filtering to estimate the current object's motion state in next frame, which contains the bounding box position information and the velocity information. More specifically, the bounding box position information and the velocity estimation of each object are shown by the orange solid box and the orange arrow in Fig. 3, respectively. Then when the false negatives fulfill the specific requirements, we apply the bounding boxes predicted by kalman filtering to complement the missing bounding boxes in the sliding window.

In particular, the first requirement is that the object must be detected by the model in the previous frame so that they can use the kalman filtering to predict the position of the object in the current frame.

Furthermore, since we employ the hungarian algorithm for frame-by-frame data association and the metric of data association is the overlap of bounding boxes, the second requirement is that the object in the previous frame must be associated by the hungarian algorithm to ensure that the tracker state of the current object is active, which can avoid adding the false positives due to sudden changes in detection results.

For the tracker that needs to look backward, the third requirement is that the object of the tracker must be detected in the last frame of the current sliding window. Since our sliding window only moves one frame at a time, as long as the detection results are correct, all false negatives that satisfy the requirements will be complemented.

The implementation details of cache track strategy are described in Algorithm 1. For each detection $\mathcal{D}_t$, if it matches the active tracker $\mathcal{T}_m$, this tracker will be updated by this detection $\mathcal{D}_t$ and a hit sign of 1 is added to the hit queue to activate the tracker. Otherwise it will initialize the new tentative tracker $T_{tentative}$, and set the first hit sign to 1. If the tentative tracker is updated by this detection $\mathcal{D}_t$, its state will change from tentative to active and a hit sign of 1 is added to its hit queue. For each tracker $\mathcal{T}_m$, if it does not match the current frame's detection $\mathcal{D}_t$, its state will change from active to tentative, and a hit sign of 0 will be added to its hit queue. If the tentative tracker is not updated for nine consecutive frames, it will be deleted from the tracker set.

In particular, the hit sign indicates whether the detector detects the object in the current frame, and we set a hit queue for each tracker to record the tracker's updated position. The length of the hit queue must not exceed the length of the sliding window. If the tracker is updated in the current frame, the hit sign of current frame in the hit queue of the tracker will be 1 otherwise 0. Therefore, the proposal algorithnm can quickly locate the false negatives for modification.

In the current sliding window, if the tracker is active and updated in the last frame of the sliding window, and the number of updates of the tracker is greater than 1 and less than

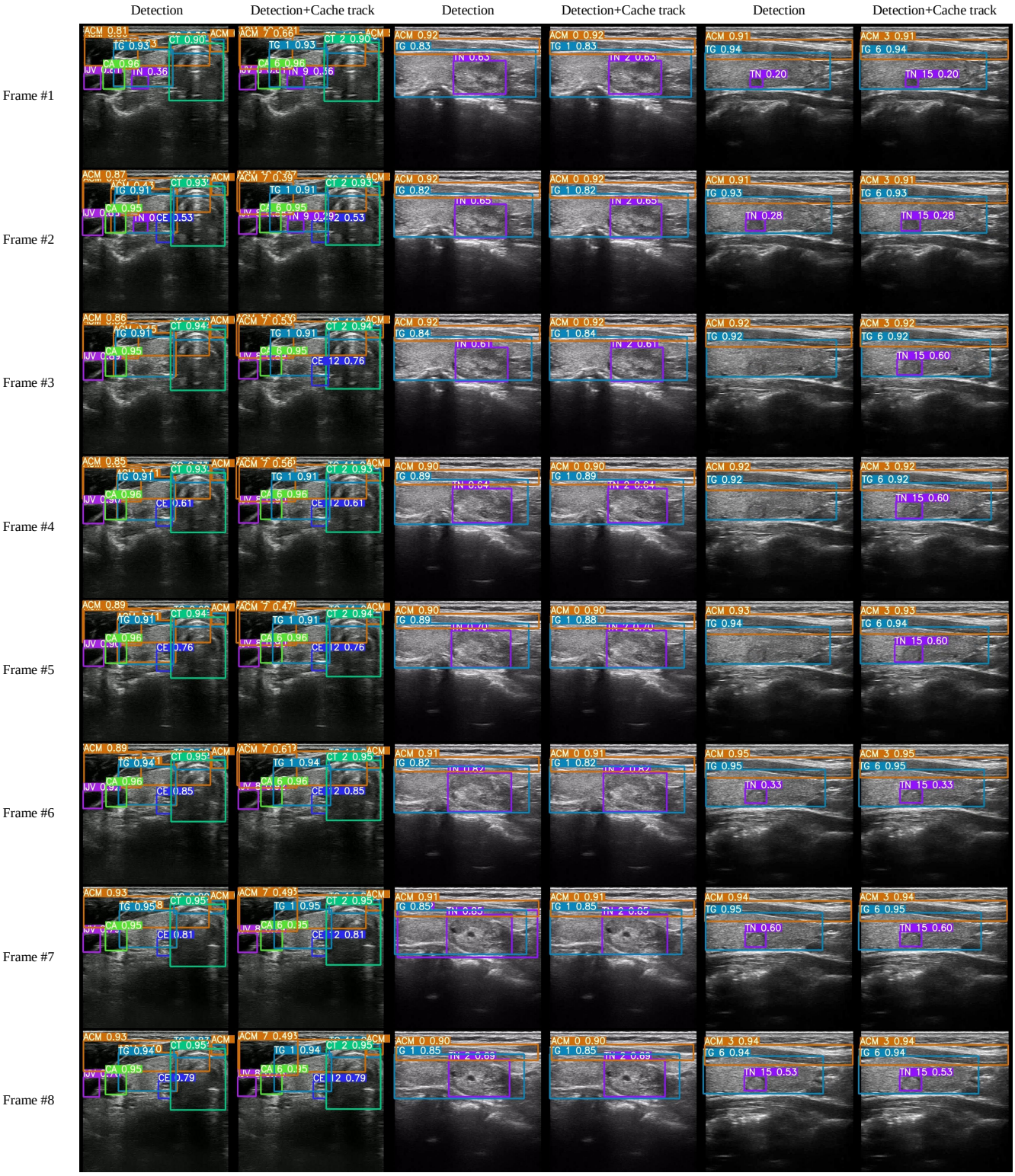

Fig. 4. Comparison results between the backbone network without cache track algorithm and the backbone network with cache track algorithm. In third row of the first two columns the missing CE is complemented by cache track algorithm. In seventh row of middle two columns the wrong TN is suppressed by cache track strategy. In third to fifth rows of last two columns the missing TN is complemented by cache track algorithm.

the length of the hit queue, the proposal algorithm will quickly locate and utilize the results predicted by kalman filtering to complement the missing bounding boxes. When the missing bounding boxes is supplemented by the prediction result of the kalman filtering, the hit sign of the current frame will be reset from 0 to 1.

Each time the sliding window will pop up the oldest frame and introduce the newest frame to keep the eight frames in the sliding window. Although we took a delay of nearly eight frames at the beginning of the video, it is still sufficiently real-time in practical application.

III. Experiments

A. Datasets and Metrics

In our study, our training videos and annotation information were obtained from the Second Affiliated Hospital and Yuying Children's Hospital of Wenzhou Medical University. The video dataset has a total of 142 thyroid ultrasound videos of 43 patients, which includes the transverse and longitudinal ultrasound scanning videos of the left and right lobes of TG. We use 130 videos of 32 patients to train the model and 12 videos of 11 patients to test.

Furthermore, to increase the credibility of the results, we also utilize another two batches of datasets for comparative experiments. The first dataset T1 contains 41 patients including 99 videos obtained from May 2020. The second dataset T2 consists of 34 videos of 17 patients acquired from June 2020. These two datasets are only used to evaluate.

Although there are about 100 frames for each video in our datasets, the number of frames for TN and CE is much less than that of other tissues. In addition, our datasets are different from the standard MOT datasets [58], and there is no associated annotation information to help the model train and evaluate the tracking performance. Since our ultrasound video datasets only contain detection annotations, and JDE, FairMot and CenterTrack do not have relevant information for learning and evaluation, we cannot make a fair comparison with these models. Therefore, we can only use the detection metrics to evaluate the performance of the model.

In order to fully verify the performance of our method, we compare not only with the same type of tracking algorithms, but also with the state-of-the-art detection models. In thyroid video detection and tracking, considering that our datasets' particularity and our method belongs to the tracking-by-detection paradigm, we utilize the detection metrics including mean average precision (mAP and mAP50) and speed (FPS) to evaluate the overall performance and average precision of each class to evaluate the performance on thyroid nodules and surrounding tissues. In precision evaluation of each class, we define that the result is accurate when the overlap between the result and the ground truth is higher than 0.5.

B. Implement details

In our experiment, we utilize the SGD optimizer to train our model. The initial learning rate of the model is 0.01, and the weight decay and the momentum are set to 5e-4 and 0.9, respectively. The batch size is 16 and the input size of the network is 640 $\times$ 640. The Data augmentation methods are implemented, which contains the mirror, random translate and color jittering, random scale and random mosaic. Specifically, we split the backbone network of darknet-53 into five stages and denote the output of each stage block as $\{C1, C2, C3, C4, C5\}$, and they have strides of $\{2, 4, 8, 16, 32\}$ pixels with respect to the input image. To improve the accuracy of small-scale objects, we leverage the feature maps of four stages $\{C2, C3, C4, C5\}$ into FPN, SPP and PAN to augment low-level image information.

In test experiment, for a fair comparison we set the confidence threshold as 0.2 and IoU threshold of NMS to be 0.65 for all methods. In tracking strategy the overlap threshold of association metric is 0.4. All the experiments in this study have been implemented on a Linux machine with the following configuration: eight Intel(R) Xeon(R) CPUs, E5-2680 v4 @ 2.40 GHZ, 256 GB of RAM, and eight NVIDIA P100 GPUs.

C. Results and discussions

Comparison on the cache track strategy. In order to verify the performance of the proposed cache track strategy, we compare our method without and with cache track algorithm. As shown in Fig. 4, we have selected three consecutive video sequences to illustrate the performance of the cache track algorithm. As presented in the first two columns of Fig. 4, it is a part of the continuous frames in the transverse ultrasound video of TG. It can be observed in the third row that when the detection model misses the bounding box of CE, the tracker interpolates the CE bounding box based on the detection results of the first two rows and the subsequent rows.

Similarly, the middle three rows of the last two columns show that the TN bounding box missed by the detection model in the longitudal ultrasound video have been successfully complemented by the cache track algorithm. These prove that our cache track algorithm can effectively complement the missing bounding box.

Additionally, as demonstrated in the seventh row of the middle two columns of Fig. 4, the detection model detects the two TN bounding boxes (purple), one of which has the same size as the TG bounding box (blue), it is clear that the larger bounding box is the wrong detection result, and the cache track strategy corrects this mistake by suppressing the wrong bounding box.

Comparison with the same type of tracking strategies. To prove the superiority of our method, we compare our method with the same type of tracking strategies including Sort [44] and DeepSort [45]. For fairness, their backbone network and parameters are the same as ours.

As presented in the lower part of Table I, Table II and Table III, the proposed cache track is better than Sort and DeepSort in terms of accuracy and comparable to them in speed. In the case that all the tracking strategies utilize the same backbone network and parameters, although our method needs time to modify the qualified detection results during the sliding window in addition to the eight-frame delay, our speed is not much slower than other methods.

Therefore, while satisfying the real-time conditions, it is worth sacrificing a bit of speed to obtain higher accuracy. In addition, it is worth noting that DeepSort performs poor especially in the detection of TN. The reason is that DeepSort [45] mainly calculates the appearance similarity as the data association metric and the appearance features of thyroid nodules are variable as illustrated in the seventh row of Fig. 5, it is difficult to associate through the appearance characteristics.

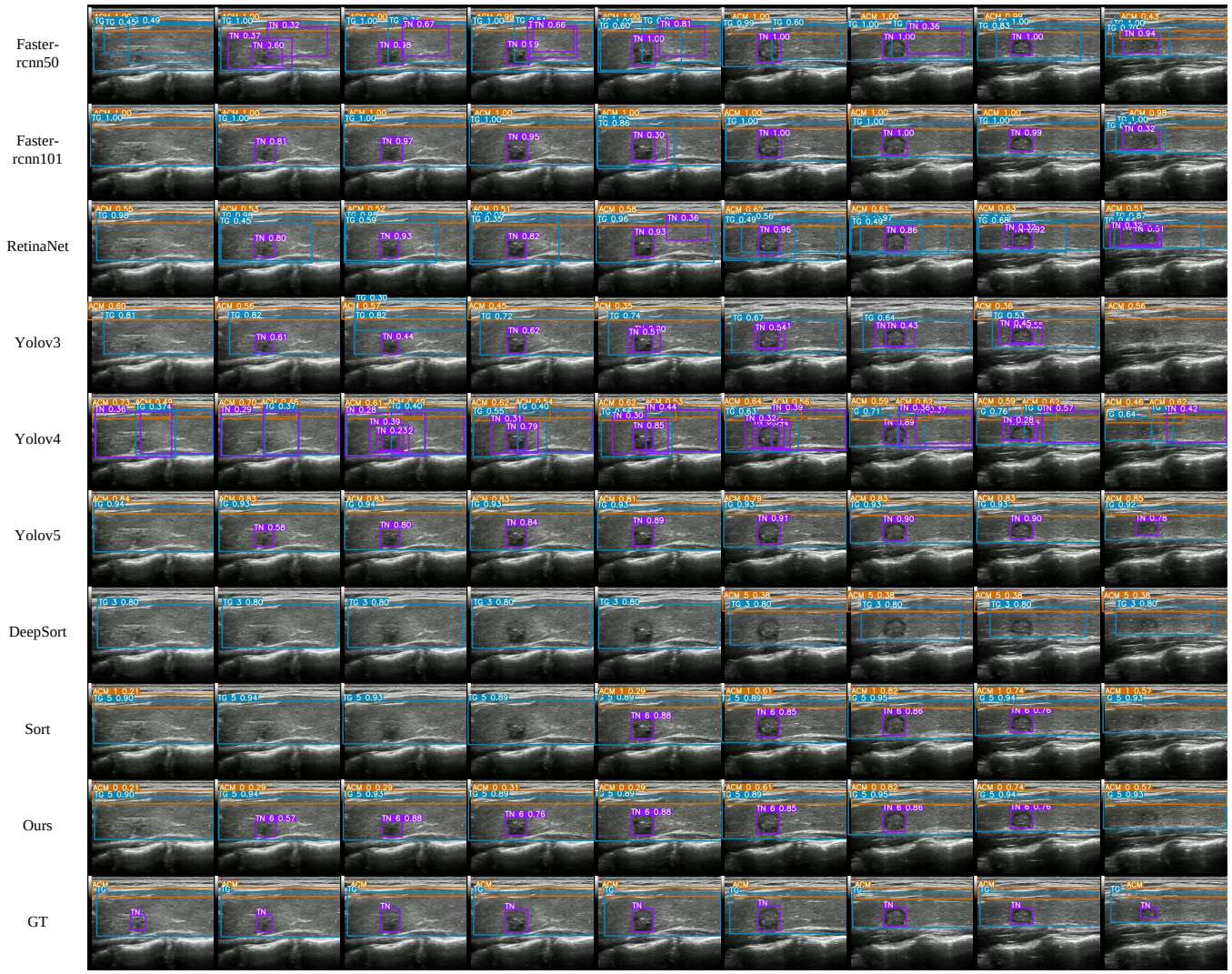

Fig. 5. Comparison results between our method and other methods. DeepSort and Sort utilize the same backbone network and parameters as ours. The ACM bounding box in the second to fourth columns and the TN in the third column in Ours row are all complemented by our cache track algorithm.

TABLE I

THE COMPARISON RESULTS BETWEEN OUR METHOD AND OTHER METHODS IN MULTI-TISSUE DETECTION AND TRACKING TASK ON TEST DATASET. IN EACH COLUMN, THE BEST RESULT IS IN **BOLD**, AND THE SECOND BEST IS UNDERLINED.

Methods	mAP	mAP50	CT	ACM	CA	IJV	TG	CE	TN	FPS
Faster RCNN [21] (Resnet50)	0.536	0.815	0.849	**0.895**	0.886	0.693	0.918	0.757	0.709	13.00
Faster RCNN [21] (Resnet101)	0.539	0.804	0.852	0.880	0.867	0.697	0.898	0.716	**0.717**	9.50
RetinaNet [27] (Resnet50)	0.501	0.786	0.885	0.695	0.885	0.723	0.896	0.771	0.646	13.10
Yolov3 [24]	0.405	0.734	0.762	0.822	0.860	0.768	0.825	0.543	0.560	30.00
Yolov4 [25]	0.555	0.871	0.960	0.823	0.939	**0.906**	0.910	0.869	0.692	**73.53**
Yolov5 [26]	**0.632**	**0.884**	**0.965**	0.863	0.953	0.895	**0.932**	0.882	0.694	67.11
DeepSort [45]	0.380	0.590	0.679	0.695	0.824	0.666	0.516	0.697	0.053	56.82
Sort [44]	0.587	0.849	0.949	0.840	0.917	0.862	0.895	0.860	0.621	57.14
Ours	0.610	0.882	**0.965**	0.857	**0.957**	0.895	0.917	**0.895**	0.688	56.18

Moreover, different from the still-image detection model, the tracking algorithm needs a probationary period (two frames) to prevent tracking of the false positives. Therefore, in the tracking algorithm, the results involved into the accuracy calculation must be the tracking results that have been successfully associated. Because the TN bounding boxes in video sequences are not matched, no TN bounding box is displayed in DeepSort row of Fig. 5 and DeepSort does not perform well on TN.

Comparison with the state-of-the-art detection methods. In order to verify that the proposed method is also comparable to the excellent detection models, we compare our method

TABLE II

THE COMPARISON RESULTS BETWEEN OUR METHOD AND OTHER METHODS IN MULTI-TISSUE DETECTION AND TRACKING TASK ON T1 DATASET. IN EACH COLUMN, THE BEST RESULT IS IN **BOLD**, AND THE SECOND BEST IS UNDERLINED.

Methods	mAP	mAP50	CT	ACM	CA	IJV	TG	CE	TN	FPS
Faster RCNN [21] (Resnet50)	0.531	0.830	0.959	0.929	0.973	0.780	0.915	0.797	0.456	12.20
Faster RCNN [21] (Resnet101)	0.536	0.830	0.960	**0.940**	0.974	0.762	0.919	0.805	0.449	9.20
RetinaNet [27] (Resnet50)	0.511	0.808	0.966	0.834	0.967	0.802	0.898	0.812	0.377	12.50
Yolov3 [24]	0.387	0.736	0.885	0.831	0.940	0.624	0.817	0.741	0.312	32.00
Yolov4 [25]	0.533	0.822	0.951	0.833	0.976	0.807	0.915	0.841	0.434	**74.63**
Yolov5 [26]	**0.555**	**0.847**	0.969	0.898	**0.986**	**0.808**	**0.927**	0.872	**0.466**	68.97
DeepSort [45]	0.363	0.621	0.791	0.708	0.893	0.640	0.604	0.640	0.069	55.25
Sort [44]	0.513	0.810	0.950	0.868	0.961	0.757	0.891	0.854	0.389	56.18
Ours	0.530	0.840	**0.971**	0.887	0.985	0.806	0.918	**0.881**	0.433	54.95

TABLE III

THE COMPARISON RESULTS BETWEEN OUR METHOD AND OTHER METHODS IN MULTI-TISSUE DETECTION AND TRACKING TASK ON T2 DATASET. IN EACH COLUMN, THE BEST RESULT IS IN **BOLD**, AND THE SECOND BEST IS UNDERLINED.

Methods	mAP	mAP50	CT	ACM	CA	IJV	TG	CE	TN	FPS
Faster RCNN [21] (Resnet50)	0.579	0.851	0.959	0.921	0.945	0.919	0.947	0.690	0.581	13.30
Faster RCNN [21] (Resnet101)	**0.594**	0.854	0.960	**0.930**	0.956	0.921	0.943	0.704	0.566	9.90
RetinaNet [27] (Resnet50)	0.554	0.829	0.965	0.820	0.944	0.923	0.921	0.710	0.518	13.00
Yolov3 [24]	0.396	0.733	0.939	0.672	0.878	0.908	0.808	0.533	0.393	29.60
Yolov4 [25]	0.533	0.812	0.956	0.628	0.949	0.928	0.928	0.733	0.560	**71.94**
Yolov5 [26]	0.593	**0.884**	0.981	0.906	0.978	0.970	**0.960**	0.734	**0.660**	70.42
DeepSort [45]	0.387	0.619	0.709	0.740	0.819	0.834	0.515	0.634	0.084	57.47
Sort [44]	0.560	0.853	0.969	0.865	0.962	0.959	0.943	0.709	0.567	56.82
Ours	0.572	0.877	**0.986**	0.871	**0.984**	**0.972**	0.953	**0.747**	0.626	56.18

with the state-of-the-art detection models (including Faster RCNN [21] with Resnet-50 [59] and Resnet-101 as backbone, RetinaNet [27] with Resnet-50 as backbone, Yolov3 [24], Yolov4 [25] and Yolov5 [26]).

For fair comparison, the datasets used to train these detection models are the same as ours. As presented in the upper part of Table I, Table II, and Table III, our method is basically comparable to the newest yolo model Yolov5 [26] and outperforms the other detection methods. Although our speed is not fastest among all methods, our method is sufficient to satisfy real-time requirements.

From the perspective of multi-tissue detection and tracking, the results of CT, CA, IJV, TG are relatively better than others and their accuracy is basically above 0.95. For ACM, Faster RCNN can achieve higher accuracy on all three test datasets. As shown in the first two Faster RCNN rows of Fig. 5, the confidence of ACM is higher than other methods, and it can be observed that the two-stage method is much better than the one-stage method in detecting the slender type of targets.

As to CE, our method performs better as our method can complement the CE bounding boxes more precisely. Unlike TN, although the number of frames that CE appears is similar to that of TN, its position and size in ultrasound video are more stable than TN. As illustrated in the first two columns of Fig. 4, our method can complement the missing CE bounding box when requirements are met.

Among all targets, TN is difficult to detect and track as its appearance, position and size are always changing in the video. Although different datasets obtain different results, our method can be roughly comparable to the state-of-the-art detection models.

As illustrated in Fig. 5, we exhibit the results of the state-of-the-art detection models and the same type of tracking strategies. We can observe that the TN in the first column is not detected by all methods due to its unclear features, and the TN in the last column is also only identified by Yolov5 and Faster RCNN.

Additionally, in order to prevent the low-confidence bounding boxes and multiple ACM bounding boxes in transverse ultrasound scanning video from being filtered out, we set the confidence threshold to 0.2 and the IoU threshold to 0.65, and it is inevitable for most still-image detection method that the abundant bounding boxes appear as shown in the first five rows of Fig. 5.

As to the tracking strategies, since we utilize the same backbone, and the detection results are same. However, the tracking results are different. As depicted in the DeepSort and Sort rows of Fig. 5, TN and ACM in the first four columns have not been tracked. For Sort, even if the ACM in the first column has been tracked before, and the ACM in the subsequent three columns is not detected, the tracker will lose three frames of ACM under the influence of the detector. However, our method can complement the missing ACM bounding boxes and TN bounding boxes as shown in the ninth row of Fig. 5.

Comparing with most state-of-the-art still-image detection models, our method is more suitable for practical application. It can not only assist the doctors in detecting and tracking the lesions and surrounding tissues, but also make statistics on the lesions and surrounding tissues. Comparing with the same type of tracking strategies, since our method can imitate the doctors' diagnosis process to look backward and modify the wrong detection results, our method is more effective and more suitable for CAD.

IV. Conclusion

In this study, we present a novel detection and tracking framework for thyroid nodules and surrounding tissues in thyroid ultrasound videos. The framework exploits the contextual information and temporal relation among video sequences as much as possible to improve the detection and tracking accuracy. Particularly, the framework is an extension of the one-stage still-image detector that introduces other important components to improve the practicability and efficiency and a new tracking strategy to solve the problem of false positives and false negatives when the still-image detector is directly applied to video detection.

To the best of our knowledge, this is the first study to detect and track thyroid nodules and surrounding tissues in thyroid ultrasound video. Experimental results show our method can achieve better performance in balancing the speed and accuracy on ultrasound video. More importantly, our method imitates the doctors' diagnostic thinking as much as possible to track and count the tissues and the lesions. We believe that this framework can help the doctors diagnose and reduce their workload. We currently aim at applying this framework into the practical application, and in the future we will work on this basis to achieve the classification of thyroid nodules in ultrasound videos.

References

[1] G. Tan, H. Chen, and J. Qi, "A novel image matting method using sparse manual clicks," *Multimedia Tools and Applications*, vol. 75, 09 2016.

[2] B. Pu, N. Zhu, K. Li, and S. Li, "Fetal cardiac cycle detection in multi-resource echocardiograms using hybrid classification framework," *Future Generation Computer Systems*, vol. 115, pp. 825 – 836, 2021. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X20307524

[3] G. Tan, Z. Guo, and Y. Xiao, *PA-RetinaNet: Path Augmented RetinaNet for Dense Object Detection*, 09 2019, pp. 138–149.

[4] X. Wu, G. Tan, K. Li, S. Li, H. Wen, X. Zhu, and W. Cai, "Deep parametric active contour model for neurofibromatosis segmentation," *Future Generation Computer Systems*, vol. 112, 05 2020.

[5] G. Tan, R. Miao, and Y. Xiao, *Action Recognition Based on Divide-and-Conquer*, 09 2019, pp. 157–167.

[6] B. Zhang, J. Tian, S. Pei, Y. Chen, X. He, Y. Dong, L. Zhang, X. Mo, W. Huang, S. Cong, and S. Zhang, "Machine learning assisted system for thyroid nodule diagnosis," *Thyroid*, vol. 29, no. 6, pp. 858–867, 2019, pMID: 30929637. [Online]. Available: https://doi.org/10.1089/thy.2018.0380

[7] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, and M. Eramian, "Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network," *Journal of Digital Imaging*, vol. 30, 07 2017.

[8] Y. Zhu, Z. Fu, and J. Fei, "An image augmentation method using convolutional network for thyroid nodule classification by transfer learning," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, 2017, pp. 1819–1823.

[9] T. Liu, S. Xie, Y. Zhang, J. Yu, L. Niu, and W. Sun, "Feature selection and thyroid nodule classification using transfer learning," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 2017, pp. 1096–1099.

[10] J. Song, Y. J. Chai, H. Masuoka, S.-W. Park, S.-j. Kim, J. Choi, H.-J. Kong, K. E. Lee, J. Lee, N. Kwak, K. Yi, and A. Miyauchi, "Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules," *Medicine*, vol. 98, p. e15133, 04 2019.

[11] R. Song, L. Zhang, C. Zhu, J. Liu, J. Yang, and T. Zhang, "Thyroid nodule ultrasound image classification through hybrid feature cropping network," *IEEE Access*, vol. PP, pp. 1–1, 03 2020.

[12] G. Shi, J. Wang, Y. Qiang, X. Yang, J. Zhao, R. Hao, W. Yang, Q. Du, and N. KAZIHISE, "Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105611, 06 2020.

[13] W. Yang, Y. Dong, Q. Du, Y. Qiang, K. Wu, J. Zhao, X. Yang, and B. Zi, "Integrate domain knowledge in training multi-task cascade deep learning model for benign–malignant thyroid nodule classification on ultrasound images," *Engineering Applications of Artificial Intelligence*, vol. 98, p. 104064, 02 2021.

[14] L. Chang, C. Fu, Z. Wu, W. Liu, and S. Yang, "Data-driven analysis of radiologists' behavior for diagnosing thyroid nodules," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, pp. 3111–3123, 2020.

[15] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *CoRR*, vol. abs/1904.07850, 2019. [Online]. Available: http://arxiv.org/abs/1904.07850

[16] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," *ArXiv*, vol. abs/1808.01244, 2018.

[17] X. Zhou, J. Zhuo, and P. Krähenbühl, "Bottom-up object detection by grouping extreme and center points," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 850–859, 2019.

[18] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9626–9635, 2019.

[19] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "Cornernet-lite: Efficient keypoint based object detection," *ArXiv*, vol. abs/1904.08900, 2020.

[20] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, 06 2015.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. Berg, "Ssd: Single shot multibox detector," in *ECCV*, 2016.

[23] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 6517–6525.

[24] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 04 2018.

[25] A. Bochkovskiy, C.-Y. Wang, and H. Liao, "Yolov4: Optimal speed and accuracy of object detection," *ArXiv*, vol. abs/2004.10934, 2020.

[26] G. Jocher, A. Stoken, J. Borovec, NanoCode012, A. Chaurasia, TaoXie, L. Changyu, A. V, Laughing, tkianai, yxNONG, A. Hogan, lorenzomammana, AlexWang1900, J. Hajek, L. Diaconu, Marc, Y. Kwon, oleg, wanghaoyang0106, Y. Defretin, A. Lohia, ml5ah, B. Milanko, B. Fineran, D. Khromov, D. Yiwei, Doug, Durgesh, and F. Ingham, "ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations," Apr. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4679653

[27] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb 2020.

[28] J. Ma, F. Wu, T. Jiang, J. Zhu, and D. Kong, "Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images," *Medical Physics*, vol. 44, 02 2017.

[29] H. Li, J. Weng, Y. Shi, W. Gu, Y. Mao, Y. Wang, W. Liu, and J. Zhang, "An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images," *Scientific Reports*, vol. 8, 12 2018.

[30] W. Song, S. Li, J. Liu, H. Qin, B. Zhang, S. Zhang, and A. Hao, "Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 1215–1224, 2019.

[31] R. Liu, S. Zhou, Y. Guo, Y. Wang, and C. Chang, "Nodule localization in thyroid ultrasound images with a joint-training convolutional neural network," *Journal of Digital Imaging*, vol. 33, 06 2020.

[32] L. Wang, S. Yang, S. Yang, C. Zhao, G. Tian, Y. Gao, Y. Chen, and Y. Lu, "Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the yolov2 neural network," *World Journal of Surgical Oncology*, vol. 17, 12 2019.

[33] V. Kumar, J. Webb, A. Gregory, D. D. Meixner, J. M. Knudsen, M. Callstrom, M. Fatemi, and A. Alizad, "Automated segmentation of thyroid nodule, gland, and cystic components from ultrasound images using deep learning," *IEEE Access*, vol. 8, pp. 63 482–63 496, 2020.

[34] D. Koundal, B. Sharma, and Y. Guo, "Intuitionistic based segmentation of thyroid nodules in ultrasound images," *Computers in Biology and Medicine*, vol. 121, p. 103776, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482520301463

[35] N. S. Narayan, P. Marziliano, J. Kanagalingam, and C. G. L. Hobbs, "Speckle patch similarity for echogenicity-based multiorgan segmentation in ultrasound images of the thyroid gland," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 172–183, 2017.

[36] J. Kang, J. Y. Lee, and Y. Yoo, "A new feature-enhanced speckle reduction method based on multiscale analysis for ultrasound b-mode imaging," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1178–1191, 2016.

[37] M. Feigin, D. Freedman, and B. W. Anthony, "A deep learning framework for single-sided sound speed inversion in medical ultrasound," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 4, pp. 1142–1151, 2020.

[38] C. Chu, J. Zheng, and Y. Zhou, "Ultrasonic thyroid nodule detection method based on u-net network," *Computer Methods and Programs in Biomedicine*, vol. 199, p. 105906, 12 2020.

[39] A. Illanes, N. Esmaeili, P. Poudel, S. Balakrishnan, and M. Friebe, "Parametrical modelling for texture characterization—a novel approach applied to ultrasound thyroid segmentation," *PLOS ONE*, vol. 14, p. e0211215, 01 2019.

[40] P. Poudel, A. Illanes, E. J. G. Ataide, N. Esmaeili, S. Balakrishnan, and M. Friebe, "Thyroid ultrasound texture classification using autoregressive features in conjunction with machine learning approaches," *IEEE Access*, vol. 7, pp. 79 354–79 365, 2019.

[41] M. Radman, M. Moradi, A. Chaibakhsh, M. Kordestani, and M. Saif, "Multi-feature fusion approach for epileptic seizure detection from eeg signals," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3533–3543, 2021.

[42] M. Kordestani, M. Saif, M. Orchard, R. Razavi-Far, and K. Khorasani, "Failure prognosis and applications–a survey of recent literature," *IEEE Transactions on Reliability*, vol. PP, 08 2019.

[43] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang, "Seq-nms for video object detection," *CoRR*, vol. abs/1602.08465, 2016. [Online]. Available: http://arxiv.org/abs/1602.08465

[44] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, pp. 3464–3468.

[45] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," 03 2017.

[46] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, and W. Ouyang, "T-cnn: Tubelets with convolutional neural networks for object detection from videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2896–2907, Oct 2018.

[47] P. Bergmann, T. Meinhardt, and L. Leal-Taixé, "Tracking without bells and whistles," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 941–951, 2019.

[48] Z. Wang, L. Zheng, Y. Liu, and S. Wang, "Towards real-time multi-object tracking," *ArXiv*, vol. abs/1909.12605, 2019.

[49] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *arXiv e-prints*, Apr. 2020.

[50] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," *ArXiv*, vol. abs/2004.01177, 2020.

[51] C.-Y. Wang, H. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "Cspnet: A new backbone that can enhance learning capability of cnn," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1571–1580, 2020.

[52] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 936–944.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 346–361.

[54] S. L. andLu Qi andHaifang Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," *CoRR*, vol. abs/1803.01534, 2018. [Online]. Available: http://arxiv.org/abs/1803.01534

[55] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[56] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," 2020.

[57] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," 2019.

[58] A. Milan, L. Leal-Taixé, I. D. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *CoRR*, vol. abs/1603.00831, 2016. [Online]. Available: http://arxiv.org/abs/1603.00831

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.