A Spatial-Temporal Deformable Attention Based Framework for Breast Lesion Detection in Videos

Chao Qin[1(✉)], Jiale Cao[2], Huazhu Fu[3], Rao Muhammad Anwer[1], and Fahad Shahbaz Khan[1,4]

[1] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
chao.qin@mbzuai.ac.ae
[2] Tianjin University, Tianjin, China
[3] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore, Singapore
[4] Linköping University, Linköping, Sweden

Abstract. Detecting breast lesion in videos is crucial for computer-aided diagnosis. Existing video-based breast lesion detection approaches typically perform temporal feature aggregation of deep backbone features based on the self-attention operation. We argue that such a strategy struggles to effectively perform deep feature aggregation and ignores the useful local information. To tackle these issues, we propose a spatial-temporal deformable attention based framework, named STNet. Our STNet introduces a spatial-temporal deformable attention module to perform local spatial-temporal feature fusion. The spatial-temporal deformable attention module enables deep feature aggregation in each stage of both encoder and decoder. To further accelerate the detection speed, we introduce an encoder feature shuffle strategy for multi-frame prediction during inference. In our encoder feature shuffle strategy, we share the backbone and encoder features, and shuffle encoder features for decoder to generate the predictions of multiple frames. The experiments on the public breast lesion ultrasound video dataset show that our STNet obtains a state-of-the-art detection performance, while operating twice as fast inference speed. The code and model are available at https://github.com/AlfredQin/STNet.

Keywords: Breast lesion detection · Ultrasound videos · Spatial-temporal deformable attention · Multi-frame prediction

1 Introduction

Ultrasound imaging is a very effective technique for breast lesion diagnosis, which has high sensitivity. Automatically detecting breast lesions is a challenging problem with a potential to aid in improving the efficiency of radiologists in ultrasound-based breast cancer diagnosis [18,21]. Some of the challenges associated with automatic breast lesion detection include blurry boundaries and changeable sizes of breast lesions.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 479–488, 2023.
https://doi.org/10.1007/978-3-031-43895-0_45

Most existing breast lesion detection methods can be categorized into image-based [10,11,16,17,19] and video-based [1,9] breast lesion detection approaches. Image-based breast lesion detection approaches perform detection in each frame independently. Compared to image-based breast lesion detection approaches, methods based on videos are capable of utilizing temporal information for improved detection performance. For instance, Chen *et al.* [1] exploited temporal coherence for semi-supervised video-based breast lesion detection. Recently, Lin *et al.* [9] proposed a feature aggregation network, termed as CVA-Net, that executes intra-video and inter-video fusions at both video and clip levels based on attention blocks. Although the recent CVA-Net aggregates clip and video level features, we distinguish two key issues that hamper its performance. First, the self-attention based cross-frame feature fusion is a global-level operation and it operates once before the encoder-decoder, thereby ignoring the useful local information and in turn missing an effective deep feature fusion. Second, CVA-Net only performs one-frame prediction based on multiple frame inputs, which is very time-consuming.

To address the aforementioned issues, we propose a spatial-temporal deformable attention based network, named STNet, for detecting the breast lesions in ultrasound videos. Within our STNet, we introduce a spatial-temporal deformable attention module to fuse multi-scale spatial-temporal information among different frames, and further integrate it into each layer of the encoder and decoder. In this way, different from the recent CVA-Net, our proposed STNet performs both deep and local feature fusion. In addition, we introduce multi-frame prediction with encoder feature shuffle operation that shares the backbone and encoder features, and only perform multi-frame prediction in the decoder. This enables us to significantly accelerate the detection speed of the proposed approach. We conduct extensive experiments on a public breast lesion ultrasound video dataset, named BLUVD-186 [9]. The experimental results validate the efficacy of our proposed STNet that has a superior detection performance. For example, our proposed STNet achieves a mAP of 40.0% with an absolute gain of 3.9% in terms of detection accuracy, while operating at two times faster, compared to the recent CVA-Net [9].

2 Method

Here, we describe our proposed spatial-temporal deformable attention based framework, named STNet, for detecting breast lesions in the ultrasound videos. Figure 1(a) presents the overall architecture of our proposed STNet, which is built on the end-to-end detector deformable DETR [22]. Within our STNet, we introduce spatial-temporal deformable attention into the encoder and the decoder. As in CVA-Net [9], we take six frames $I_{k-1}, I_k,\ I_{k+1},\ I_{r1},\ I_{r2},\ I_{r3}$ from one ultrasound video as inputs, where there are three neighboring frames $I_{k-1}, I_k,\ I_{k+1}$ and three randomly-selected frames $I_{r1},\ I_{r2},\ I_{r3}$. Given these input frames, we use the backbone, such as ResNet-50 [6], to extract deep multi-scale features $F_{k-1}, F_k,\ F_{k+1},\ F_{r1},\ F_{r2},\ F_{r3}$. Afterwards, we introduce a

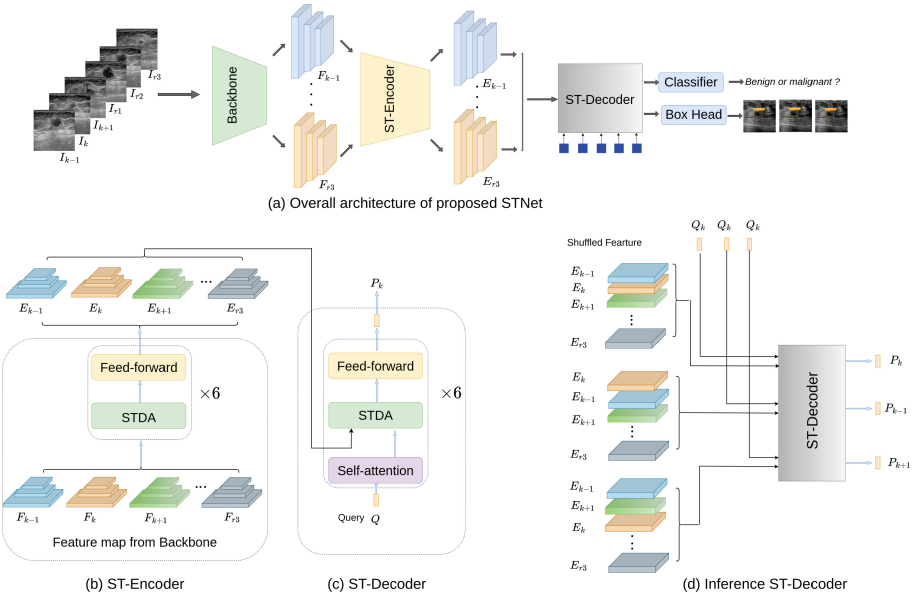

Fig. 1. (a) Overall architecture of the proposed STNet. The proposed STNet takes six frames as inputs and extracts multi-scale features of each frame. Afterwards, the proposed STNet utilizes a spatial-temporal deformable attention (STDA) based encoder **(b)** and decoder **(c)** for spatial-temporal multi-scale information fusion. Finally, the proposed STNet performs classification and regression. **(d)** During inference, we introduce a encoder feature shuffle strategy for multi-frame prediction.

spatial-temporal deformable attention based encoder (ST-Encoder) to perform intra-frame and inter-frame multi-scale feature fusion. Then, we introduce a spatial-temporal deformable attention based decoder (ST-Decoder) to generate output feature embeddings P_k, which are fed to a classifier and a box predictor for classification and bounding-box regression. During inference, we take three neighboring frames and three randomly-selected frames as the inputs, and simultaneously predict the results of three neighboring frames using our encoder feature shuffle strategy. As a result, our approach operates at a faster inference speed.

2.1 Spatial-Temporal Deformable Attention

Given a reference point, deformable attention [22] aggregates the features of a group of key sampling points near it. Compared to original transformer self-attention [13], deformable attention has low-complexity along with a faster convergence speed. Motivated by this, we adopt deformable attention for breast lesion detection and extend it to spatial-temporal deformable attention (STDA). Our STDA not only aggregates the features of current frame, but also aggregates the features of the rest of the frames. Figure 2 presents the structure of our

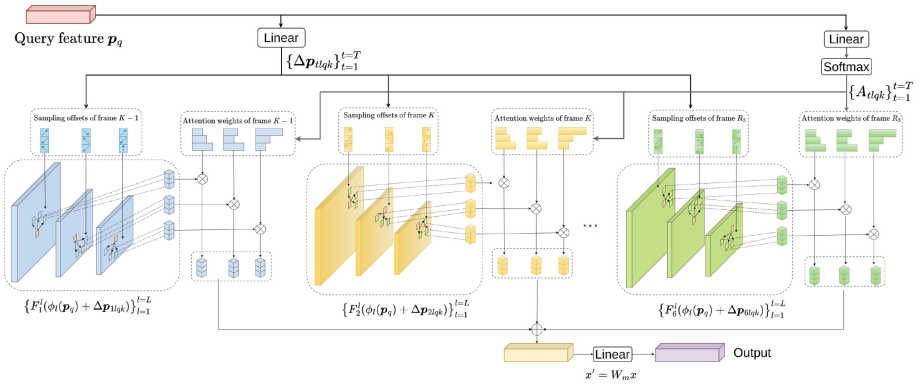

Fig. 2. Structure of our proposed Spatial-temporal deformable attention (STDA). Given a query feature and reference point, our STDA not only fuses multi-scale features within a frame, but also aggregates multi-scale features between different frames.

proposed STDA. Let $F_t = \left\{F_t^l\right\}_{l=1}^{L}$ represent the set of multi-scale feature maps at frame t, where $F_t^l \in \mathbb{R}^{C \times H_l \times W_l}$ is the feature map at level l. Given the query features $\boldsymbol{p}_q$ and corresponding reference points $\boldsymbol{z}_q$, the spatio-temporal multi-scale attention is given as:

$$\text{STDA}\left(\boldsymbol{z}_q, \boldsymbol{p}_q, \{F_t\}_{t=0}^{T}\right) = \sum_{m=1}^{M} W_m \sum_{t=1}^{T}\sum_{l=1}^{L}\sum_{k=1}^{K} A_{tlqk} F_t^l(\phi_l(\boldsymbol{p}_q) + \Delta \boldsymbol{p}_{tlqk}), \quad (1)$$

where m represents multi-head index and k is sampling point index. W_m is a linear layer, A_{tlqk} indicates attention weight of sampling point, and $\Delta \boldsymbol{p}_{tlqk}$ indicates sample offset of sampling point. ϕ_l normalizes the coordinates $\boldsymbol{p}_q$ by the scale of feature map F_t^l. The sampling offset $\Delta \boldsymbol{p}_{tlqk}$ is predicted by the query feature $\boldsymbol{z}_q$ with a linear layer. The attention weight A_{tlqk} is predicted by feeding query feature $\boldsymbol{z}_q$ to a linear layer and a softmax layer. As a result, the sum of attention weights is equal to one as

$$\sum_{t=1}^{T}\sum_{l=1}^{L}\sum_{k=1}^{K} A_{tlqk} = 1. \quad (2)$$

Compared to the standard deformable attention, the proposed spatial-temporal deformable attention fully exploits spatial information within frame and temporal information across frames.

2.2 Spatial-Temporal Deformable Attention Based Encoder and Decoder

Here, we integrate the proposed spatial-temporal deformable attention (STDA) into encoder and decoder (called ST-Encoder and ST-Decoder). As shown in

Fig. 1(b), ST-Encoder takes deep multi-scale feature maps $F_{k-1}, F_k, F_{k+1}, F_{r1}, F_{r2}, F_{r3}$ as inputs. Afterwards, we employ STDA to perform spatial and temporal fusion and generate the fused multi-scale feature maps $F'_{k-1}, F'_k, F'_{k+1}, F'_{r1}, F'_{r2}, F'_{r3}$, where the query corresponds to each pixel in multi-scale feature maps. Then, the fused feature map goes through a feed-forward network (FFN) to generate the output feature maps $E_{k-1}, E_k, E_{k+1}, E_{r1}, E_{r2}, E_{r3}$. Similar to the original deformable DETR, we adopt cascade structure to stack six STDA and FFN layers in ST-Encoder.

The ST-Decoder takes the output feature maps $E_{k-1}, E_k, E_{k+1}, E_{r1}, E_{r2}, E_{r3}$ and a set of learnable queries $Q \in \mathbb{R}^{N \times C}$ as inputs. The learnable queries first go through a self-attention layer. Afterwards, STDA performs cross-attention operation between these feature maps and the queries, where the key elements are these output feature maps of ST-Encoder. Then, we employ a FFN layer to generate the prediction features $P_k \in \mathbb{R}^{N \times C}$. We also stack six self-attention, STDA, and FFN layers in ST-Decoder for deep feature extraction.

2.3 Multi-frame Prediction with Encoder Feature Shuffle

As discussed above, the proposed STNet adopts six frames to predict the results of one frame. Although STNet fully exploits temporal information for improved breast lesion detection, it becomes time-consuming for multi-frame prediction. To accelerate the detection speed, we introduce multi-frame prediction with encoder feature shuffle during inference. Instead of going through the entire network several times, we first share deep multi-scale feature maps before encoder and second perform the decoder several times for multi-frame prediction. To perform multi-frame prediction only in the decoder, we propose the encoder feature shuffle operation shown in Fig. 1(d). By exchanging the order of neighboring frame I_{k-1}, I_k, I_{k+1}, the decoder can predict the results of three neighboring frames, respectively. Compared to the original STNet, the proposed encoder feature shuffle strategy only employs decoder forward three frames and accelerates the inference speed.

3 Experiments

3.1 Dataset and Implementation Details

Dataset. We conduct the experiments on the public BLUVD-186 dataset [9], comprising 186 videos including 112 malignant and 74 benign cases. The dataset has totally 25,458 ultrasound frames, where the number of frames in a video ranges from 28 to 413. The videos encompass a comprehensive tumor scan, from its initial appearance to its largest section and eventual disappearance. All videos were captured using PHILIPS TIS L9-3 and LOGIQ-E9. The grounding-truths in a frame, including breast lesion bounding-boxes and corresponding categories, are labeled by two pathologists, which have eight years of professional background in the field of breast pathology. We adopt the same dataset splits as in

Table 1. State-of-the-art quantitative comparison of our approach with existing methods in literature on the BLUVD-186 dataset. Our approach achieves a superior performance on three different metrics. Compared to the recent CVA-Net [9], our approach obtains a gain of 3.9% in terms of overall AP. We show the best results in bold.

Method	Type	Backbone	AP	AP_{50}	AP_{75}
GFL [7]	image	ResNet-50	23.4	46.3	22.2
Cascade RPN [14]	image	ResNet-50	24.8	42.4	27.3
Faster R-CNN [12]	image	ResNet-50	25.2	49.2	22.3
VFNet [20]	image	ResNet-50	28.0	47.1	31.0
RetinaNet [8]	image	ResNet-50	29.5	50.4	32.4
DFF [24]	video	ResNet-50	25.8	48.5	25.1
FGFA [23]	video	ResNet-50	26.1	49.7	27.0
SELSA [15]	video	ResNet-50	26.4	45.6	29.6
Temporal ROI Align [5]	video	ResNet-50	29.0	49.9	33.1
MEGA [2]	video	ResNet-50	32.3	57.2	35.7
CVA-Net [9]	video	ResNet-50	36.1	65.1	38.5
STNet (Ours)	video	ResNet-50	**40.0**	**70.3**	**43.3**

the previous work CVA-Net [9], to guarantee a fair comparison. Specifically, the testing set comprises 38 videos randomly selected from all 186 videos, while the rest of the videos are used as the training set.

Evaluation Metrics. Three commonly-used metrics are employed for performance evaluation of breast lesion detection methods on the ultrasound videos, namely average precision (AP), AP_{50}, and AP_{75}.

Implementation Details. We employ the ResNet-50 [6] pre-trained on ImageNet [3], and use Xavier [4] to initialize the remaining network parameters. To enhance the diversity of training data, all videos are randomly subjected to horizontal flipping, cropping, and resizing. Similar to that of CVA-Net, we employ a two-phase training strategy to achieve better convergence. In the first phase, we employ Adam optimizer to train the model for 8 epochs. We then fine-tune the model for another 20 epochs with the SGD optimizer. Throughout both phases of training, we adopt the consistent hyper-parameters, where the learning rate is 5×10^{-5} and the weight decay is 1×10^{-4}. We train the model on a single NVIDIA A100 GPU and set the batch size as 1.

3.2 State-of-the-Art Comparison

Our proposed approach is compared with eleven state-of-the-art methods, comprising image-based and video-based methods. We report the detection performance of these state-of-the-art methods generated by CVA-Net [9]. Specifically, CVA-Net acquires the detection performance of these methods by utilizing their publicly available codes or re-implementing them if no publicly available codes.

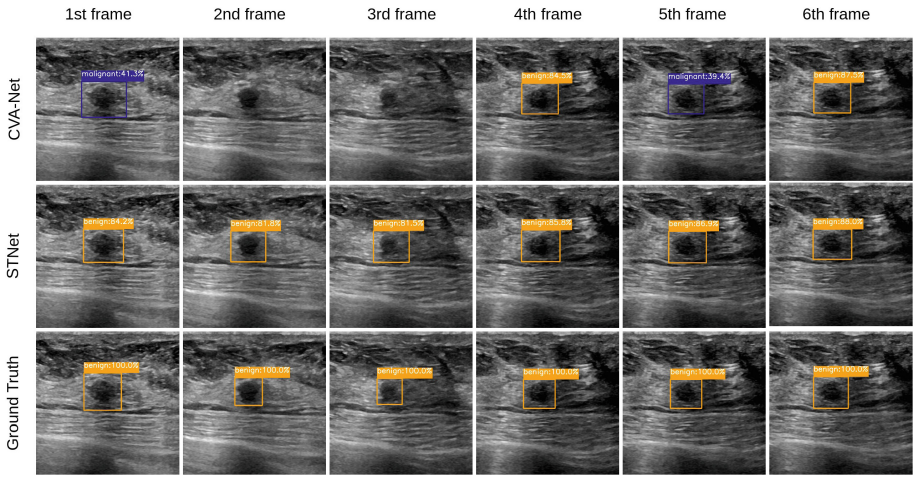

Fig. 3. Qualitative breast lesion detection comparison on example ultrasound video frames between the recent CVA-Net [9] and our proposed STNet. We also show the ground truth as reference. Our STNet achieves improved detection performance, compared to CVA-Net. Best viewed zoomed in. (Color figure online)

Quantitative Comparisons. Table 1 presents the state-of-the-art quantitative comparison of our approach with the eleven existing breast lesion video detection methods in literature. As a general trend, video-based methods tend to yield higher average precision (AP), AP50, and AP75 scores compared to image-based breast lesion detection methods. Among the eleven existing methods, the recent CVA-Net [9] achieves the best overall AP score of 36.1, AP50 score of 65.1, and AP75 score of 38.5. Our proposed STNet method consistently outperforms CVA-Net [9] on all three metrics (AP, AP50, and AP75). Specifically, our STNet achieves a significant improvement in the overall AP score from 36.1 to 40.0, the AP50 score from 65.1 to 70.3, and the AP75 score from 38.5 to 43.3. The significant improvement demonstrates the efficacy of our approach for detecting breast lesions in ultrasound videos.

Qualitative Comparisons. Figure 3 presents the qualitative breast lesion detection comparison between CVA-Net and our proposed approach on an ultrasound video containing the benign breast lesions. Moreover, we show the ground truth of each frame on the third row for reference. The first row of the figure shows that CVA-Net struggles to identify the breast lesions in the second and third frames. Further, although CVA-Net manages to identify the breast lesions in the first and fifth frames, the classification results are inaccurate (as highlighted by the blue rectangle in Fig. 3). In contrast, our STNet method in the second row of Fig. 3 accurately detects the breast lesions in all video frames and achieves accurate classification performance for each frame.

Table 2. Ablation study with different design choices. Our proposed STNet achieves a superior performance compared to the baseline and some different designs. We show the est results in bold.

	AP	AP_{50}	AP_{75}
Baseline + Single-frame	30.2	55.0	31.7
Baseline + Multi-frame	35.1	61.6	37.4
ST-Encoder + DA-Decoder	34.9	59.8	37.7
DA-Encoder + ST-Decoder	35.8	60.4	38.0
STNet (Ours)	**40.0**	**70.3**	**43.3**

Inference Speed Comparison. We present the inference speed comparison between our proposed STNet and CVA-Net on an NVIDIA RTX 3090 GPU using the same environment. We use FPS (frames per second) as the performance metric. Specifically, our proposed STNet achieves an averaged inference speed of 21.84 FPS, while CVA-Net achieves an averaged speed of 12.17 FPS. Our model operates around two times faster than CVA-Net, which we attribute to the ability of our model to predict three frames simultaneously.

3.3 Ablation Study

Effectiveness of STDA: To show the efficacy of our proposed STDA, we perform different ablation studies. The first baseline network, referred as "Baseline + Single-frame", uses the original deformable DETR and takes a single frame as input. The second baseline network, referred as "Baseline + Multi-frame", uses modified deformable DETR with multi-head attention module to fuse six input frames. For the third study, labeled "ST-Encoder + DA-Decoder", we retain the encoder with STDA in our model but replace the STDA in the decoder with the conventional deformable attention. Similarly, in the fourth study, labeled "DA-Encoder + ST-Decoder", we retain the decoder with STDA in our model but replace the STDA in the encoder with the conventional deformable attention. As shown in Table 2, the results show that "ST-Encoder + DA-Decoder" and "DA-Encoder + ST-Decoder" improve the AP by 4.7 and 5.6, respectively, compared to "Baseline + Single-frame". This demonstrates that STDA can effectively perform intra-frame and inter-frame multi-scale feature fusion, even when only partially adopted in the encoder or decoder. Furthermore, our proposed STNet improves the AP by 5.1 and 4.2 compared to "ST-Encoder + DA-Decoder" and "DA-Encoder + ST-Decoder", respectively, indicating that the integration of STDA in both the encoder and decoder is crucial for achieving superior detection performance.

4 Conclusion

We propose a novel breast lesion detection approach for ultrasound videos, termed as STNet, which performs local spatial-temporal feature fusion and

deep feature aggregation in each stage of both encoder and decoder using our spatial-temporal deformable attention module. Additionally, we introduce the encoder feature shuffle strategy that enables multi-frame prediction during inference, thereby enabling us to accelerate the inference speed while maintaining better detection performance. The experiments conducted on a public breast lesion ultrasound video dataset show the efficacy of our STNet, resulting in a superior detection performance while operating at a fast inference speed. We believe STNet presents a promising solution and will help further promote future research in the direction of efficient and accurate breast lesion detection in videos.

Acknowledgment. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-TC-2021-003) and Agency for Science, Technology and Research (A*STAR) Central Research Fund (CRF).

References

1. Chen, S., et al.: Semi-supervised breast lesion detection in ultrasound video based on temporal coherence. arXiv:1907.06941 (2019)
2. Chen, Y., Cao, Y., Hu, H., Wang, L.: Memory enhanced global-local aggregation for video object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10337–10346 (2020)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
4. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
5. Gong, T., et al.: Temporal Rol align for video object recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1442–1450 (2021)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
7. Li, X., et al.: Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. Adv. Neural. Inf. Process. Syst. **33**, 21002–21012 (2020)
8. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
9. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., Wang, L.: A new dataset and a baseline model for breast lesion detection in ultrasound videos. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. LNCS, vol. 13433. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_59
10. Movahedi, M.M., Zamani, A., Parsaei, H., Tavakoli Golpaygani, A., Haghighi Poya, M.R.: Automated analysis of ultrasound videos for detection of breast lesions. Middle East J. Cancer **11**(1), 80–90 (2020)
11. Qi, X., et al.: Automated diagnosis of breast ultrasonography images using deep neural networks. Med. Image Anal. **52**, 185–198 (2019)

12. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497 (2015)
13. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
14. Vu, T., Jang, H., Pham, T.X., Yoo, C.D.: Cascade RPN: delving into high-quality region proposal network with adaptive convolution. arXiv preprint arXiv:1909.06720 (2019)
15. Wu, H., Chen, Y., Wang, N., Zhang, Z.: Sequence level semantics aggregation for video object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9217–9225 (2019)
16. Xue, C., et al.: Global guidance network for breast lesion segmentation in ultrasound images. Med. Image Anal. **70**, 101989 (2021)
17. Yang, Z., Gong, X., Guo, Y., Liu, W.: A temporal sequence dual-branch network for classifying hybrid ultrasound data of breast cancer. IEEE Access **8**, 82688–82699 (2020)
18. Yap, M.H., et al.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J. Biomed. Health Inform. **22**(4), 1218–1226 (2017)
19. Zhang, E., Seiler, S., Chen, M., Lu, W., Gu, X.: BIRADS features-oriented semi-supervised deep learning for breast ultrasound computer-aided diagnosis. Phys. Med. Biol. **65**(12), 125005 (2020)
20. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: VarifocalNet: an IoU-aware dense object detector. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8510–8519 (2021)
21. Zhu, L., et al.: A second-order subregion pooling network for breast lesion segmentation in ultrasound. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 160–170. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_16
22. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)
23. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 408–417 (2017)
24. Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y.: Deep feature flow for video recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2349–2358 (2017)