Medical Image Analysis 80 (2022) 102490

Contents lists available at ScienceDirect

Medical Image Analysis

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

Extracting keyframes of breast ultrasound video using deep reinforcement learning

Ruobing Huang[a], Qilong Ying[a], Zehui Lin[a], Zijie Zheng[b], Long Tan[b], Guoxue Tang[b], Qi Zhang[b], Man Luo[b], Xiuwen Yi[b], Pan Liu[b], Weiwei Pan[c], Jiayi Wu[b,*], Baoming Luo[b,*], Dong Ni[a,*]

[a] Medical UltraSound Image Computing (MUSIC) Lab, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China
[b] Department of Ultrasound, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China
[c] Department of Ultrasound, Junan Hospital Affiliated to Shunde Hospital of Guangzhou University of Chinese Medicine (Foshan Shunde Junan Hospital), Foshan, China

ARTICLE INFO

Article history:
Received 24 January 2022
Revised 8 April 2022
Accepted 20 May 2022
Available online 5 June 2022

Keywords:
Ultrasound
Breast cancer
Reinforcement learning
Video summarization
Keyframe extraction

ABSTRACT

Ultrasound (US) plays a vital role in breast cancer screening, especially for women with dense breasts. Common practice requires a sonographer to recognize key diagnostic features of a lesion and record a single or several representative frames during the dynamic scanning before performing the diagnosis. However, existing computer-aided diagnosis tools often focus on the final diagnosis process while neglecting the influence of the keyframe selection. Moreover, the lesions could have highly-irregular shapes, varying sizes, and locations during the scanning. The recognition of diagnostic characteristics associated with the lesions is challenging and also faces severe class imbalance. To address these, we proposed a reinforcement learning-based framework that can automatically extract keyframes from breast US videos of unfixed length. It is equipped with a detection-based nodule filtering module and a novel reward mechanism that can integrate anatomical and diagnostic features of the lesions into keyframe searching. A simple yet effective loss function was also designed to alleviate the class imbalance issue. Extensive experiments illustrate that the proposed framework can benefit from both innovations and is able to generate representative keyframe sequences in various screening conditions.

1. Introduction

According to data released by World Health Organization in 2020, female breast cancer is the most commonly occurring cancer worldwide (11.7% of the total new cases) (Organization et al., 2020). Early diagnosis of breast cancer opens the door to timely treatment and long-term survival rate (Senie et al., 1981). Being non-invasive, non-ionizing, and cost-effective, ultrasound (US) imaging plays a vital role in breast cancer diagnosis (Shen et al., 2015), especially for women with dense breasts. However, accurately recognizing and diagnosing the lesions require expertise and clinical experience. The Breast Imaging-Reporting and Data System (BI-RADS) (American College of Radiology and others, 2003) was published as a clinical guideline to standardize the diagnosis process. It defines a series of ultrasonic characteristics, including the edge, shape, orientation, etc., which are critical indicators of malignancy that the sonographer should recognize. In practice, the sonographer moves the probe to identify representative frames of a lesion where the most (or the clearest) of these indicators can be visualized. These frames are subsequently saved and the final diagnosis is made based on them. This process is indispensable, as some diagnostic attributes could only be visualized in certain cross-sections of a lesion. For example, in the first row of Fig. 1, the lesion presents a circular shape in the first frame (upper-left corner). However, its shape changes dramatically to a highly irregular one with a slight tilt of the US probe (i.e., the third frame), which indicates a higher risk of malignancy. Similarly, the fifth row of Fig. 1 exhibits a case where calcification and the irregularity of the lesion shape can be better observed in the fifth frame, while this information may elude in other frames. Therefore, an inappropriate selection of frames might lead to incorrect interpretation of the attributes of lesions and result in misdiagnosis. An automated approach is desired to assist the less-experienced sonographers to identify suitable frames during diagnosis.

Over the years, many inspiring works have been published on classifying the lesions using static US images (Han et al., 2017;

* Corresponding authors.
E-mail addresses: wujiayi@mail2.sysu.edu.cn (J. Wu), luobm@mail.sysu.edu.cn (B. Luo), nidong@szu.edu.cn (D. Ni).

https://doi.org/10.1016/j.media.2022.102490
1361-8415/

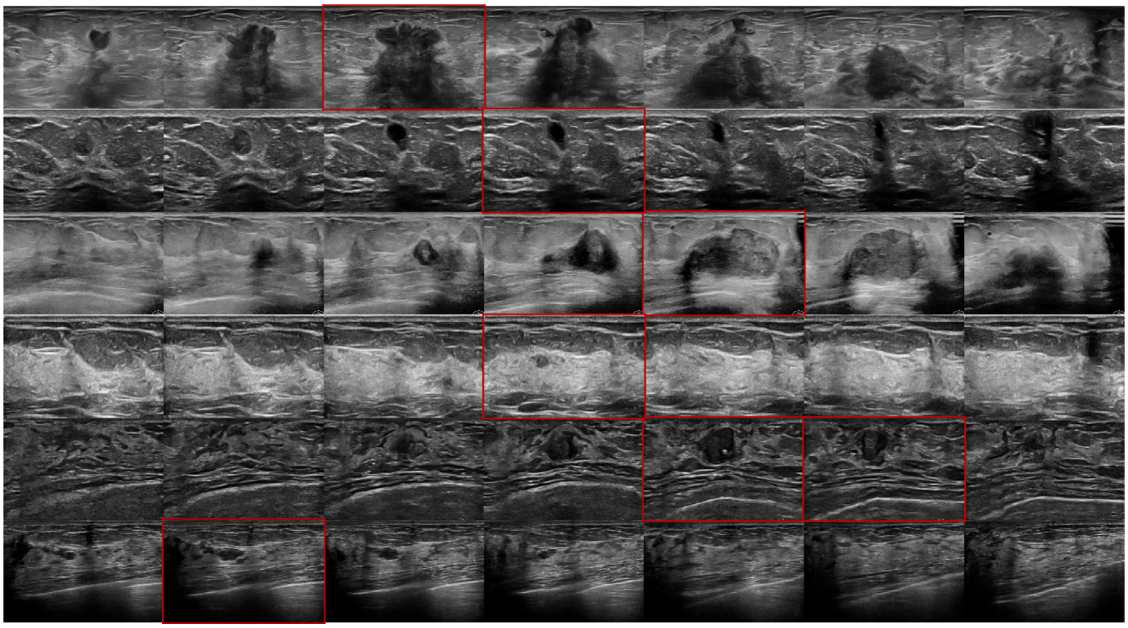

Fig. 1. Examples of breast US videos. Lesions with varying sizes and shapes appear at different time points and spatial locations within the videos (e.g. compare row 1 to 2, row 3 to 4). Representative frames of each lesion are highlighted in red. Key diagnostic attributes can be visualized in some of the frames while this information eludes other frames (e.g. calcification in row 5, micro-lobulated margin in row 1). The overall contrast and brightness of US videos are also different (e.g. compare row 4 to 6), adding further difficulty during data analysis.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Mohammed et al., 2018; Ting et al., 2019; Wang et al., 2020; Huang et al., 2021). However, to the best of our knowledge, identifying which representative static image to classify within a full breast US scan was mainly performed manually by the sonographers during data collection. The accuracy and reproducibility of this process have been generally overlooked. Nevertheless, this task can be prompted by existing approaches in keyframe extraction and video summary, both have been explored in other applications. The video summary technique finds a brief and diversified representation of the content (Parihar et al., 2021) and can facilitate the browsing (Rahman et al., 2020), and retrieval (Huang and Worring, 2020) of videos. It usually summarizes the whole video into an unfixed amount of clips (typically taking up 15% of the original ones) (Otani et al., 2019; Pan et al., 2019; Liu et al., 2020a; Apostolidis et al., 2021). Keyframe extraction, on the other hand, also has a wide range of applications. In the field of medical image analysis, it was mainly used to find a fixed amount of frames/planes within a video/volume to represent certain anatomical structures (Ciompi et al., 2011; Baumgartner et al., 2016; Chen et al., 2017; Dou et al., 2019).

Note that finding keyframes of breast lesions in US videos is slightly different from both. It requires a more condensed representation than that provided by a video summary while the number of frames is not pre-fixed. Moreover, existing applications usually endeavor to find keyframes defined by certain anatomy with a unique shape or spatial location (e.g., trans-cerebellar plane and abdominal plane in fetal sonography, four-chamber view in echo-cardiography). This is inherently different from that used in breast US scans as breast lesions cannot be uniformly characterized. Moreover, this mission faces the following challenges:

(1) There exists a large amount of redundant information as the lesion might only be visible within a small portion of the whole scan (e.g. row 4 and 6 in Fig. 1). The proposed model should only provide keyframes of the lesion and is not distracted by misleading surrounding tissues.

(2) The appearance of the lesions and their keyframes can be drastically different. The location, size, and shape of lesions have large variations between different patients (e.g., compare row 3 and 4 in Fig. 1), and even across the same US scan. As shown in the first row of Fig. 1, the size of the lesion changes drastically from frame 1 to 5. As a result, the keyframes of each lesion could only be determined based on its own characteristics and might be prone to subjectivity.

(3) The need of incorporating clinical prior knowledge within the model design. Following the previous point, the keyframes of the lesions have no ubiquitous features while their definition should be correlated and beneficial to the diagnosis. Therefore, it is desired to incorporate diagnostic-related information within the frame selection process, which can also promote the interpretability of the model.

To address these challenges, we design a novel framework by simulating the behavior of sonographers during a breast US examination. The main contributions are:

- A reinforcement learning-based framework that could dynamically output a keyframe set to represent each unique lesion. It is equipped with a customized reward function that integrates guidance from manual annotation, nodule presence, and diagnostic attributes. This design also amplifies the model interpretability with clinically understandable terms.
- A detection-based nodule filtering module that can effectively remove redundant information and auto-focus the search to the lesion region. It also allows the proposed framework to process video with an unfixed length.
- A novel group-aware focal loss for accurate multi-attribute classification. It is designated to handle class imbalance through decoupling the optimization between the majority group and the minority one.

2. Related works

In this era of information explosion, videos have become ubiquitous in everyone's daily life. Due to the sheer amount and volume of data contains in these videos, it is often infeasible to analyze the whole archived dataset without the help of modern computer science techniques. Plenty of methods have been proposed

to compress the videos into a short, manageable summary or extract the most relevant information through selecting representative keyframes. Note also that these two terms–video summarization and keyframe detection–are used interchangeably in some of the literature due to the close affinity between the two tasks. Here we focus the discussion on the more recent deep learning-based methods proposed for analyzing videos.

2.1. Video summarization

Classical approaches usually first extract frame-level features based on pre-trained CNN models, then use a recurrent model (e.g., LSTM) to capture the temporal dependency (Zhang et al., 2016). Mahasseni et al. (2017) followed this design, and they introduced an additional LSTM model to minimize the distance between the feature of the whole video and that of the summary in an unsupervised way. Rochan et al. (2018) posed video summarization as a semantic segmentation task and leveraged fully convolutional networks to produce a binary mask where each element indicates whether the frame is selected for the summary. Fajtl et al. (2018) also proposed to discard the recurrent network, while they replaced it with the attention mechanism, which is also attached with a fully-connected layer to regress the frame importance score. Zhou et al., 2018 also used a pre-trained CNN to extract features of video frames, while they developed a deep summarization network based on the RL technique to jointly account for diversity and representativeness of generated summaries. Following this work, Liu et al. (2020a) applied a similar method to fetal ultrasound videos to predict the likelihood of a frame being a standard diagnostic plane. These standard planes are defined by the presence of typical anatomical structures such as the brain, lips, kidneys, etc. Note that these methods mostly deal with datasets with dense annotations as many or even all of the frames are labeled with different importance scores. Conversely, our dataset only has one annotated keyframe for each video, which leads to an extremely sparse supervision signal during training. Moreover, existing approaches usually use separate models for frame feature extraction and summarization generation, while our method train the two end-to-end and allows the model to learn bespoke features for video summarization.

2.2. Keyframe detection

Keyframe detection can facilitate indexing, classification, and further evaluation of high-dimensional data and has a wide range of applications in both computer vision and medical image analysis. Here, we mainly discuss literature that is related to US imaging, while other techniques could be referred to in the review papers, such as Asha Paul et al. (2018).

Some of the earlier approaches rely on morphological features to detect keyframes in US sequences (Ciompi et al., 2011). The modern technique leverages the power of neural networks to automatically extract features. For example, Baumgartner et al. (2016) and Stoean et al. (2021) both formulated the keyframe detection as a classification problem. The former used a 6-layer CNN to classify whether a frame belongs to 12 standard planes in the fetal US, while the latter targeted at classifying four keyframes of the fetal heart. Chen et al. (2017) combined specialized CNN and RNN to exploit the in- and between-plane features to detect three standard planes of fetus US videos (i.e., face, chest, and abdomen).

Other methods proposed to combine information from other resources to further boost detection accuracy. Cai et al. (2020) leveraged the tracked gaze of sonographers to generate visual attention maps and further improve keyframe detection accuracy in fetal US videos. Jiao et al. (2020) incorporated speech data to learn cross-modal video-speech representations in a self-supervised manner, which was demonstrated to be effective for the downstream standard plane detection task, also in fetal US scans. Pu et al. (2021) combined the merits of methods such as Yan et al. (2018) and Chen et al. (2017) by first extracting CNN features from both the US sequence and the optical flow, then they used RNN to fuse their features for standard plane recognition. These approaches are targeted at fetal US videos of different organs, whose keyframes are defined by the presence of anatomical structures with similar appearances. Breast lesions, however, lack unity in shapes or sizes. Therefore, it may be critical to incorporate additional clinical knowledge to further guide the frame extraction process.

2.3. Class imbalance

An additional challenge this work faces is the class imbalance issue encountered in lesion attribute recognition. This is essentially a common problem for many machine learning applications when there exists a majority class that dominates the dataset and leads to biased predictions (Buda et al., 2018). A classical solution to this is the sampling-based technique, which includes under-sampling, over-sampling, and a hybrid of both (Yap et al., 2014). However, they could either cause information loss or repetitive computation. Experimental results have also shown that the performance of such methods is highly dependent on both the learner and the evaluation metrics (Van Hulse et al., 2007). Moreover, as different nodule attributes exhibit different majority to minority ratios (see Table 1), it is infeasible to find a sampling strategy that can balance all the distributions.

Cost function-based methods, on the other hand, do not require manipulation of data and could be applied to different models. For example, Wang et al., 2016 proposed an MFE loss where they modified the traditional mean squared error loss by calculating the error of each class separately. The Focal loss proposed by Lin et al. (2017) is another popular choice. They added an additional factor $(1-p)^{\gamma}$ to the standard cross-entropy loss to reduce the effects of easy examples. Recently, Li et al. (2019) proposed a gradient harmonizing mechanism to obtain a balance between the positive and negative examples. Wu et al. (2020) introduced a re-balancing weight to tackle label co-occurrence. Ridnik et al. (2021) proposed the ASL loss that decouples the focusing level of the positive and negative samples with shifted probability. We opt for the cost function-based approach to address the imbalanced data in attribute classification.

3. Methodology

Interpreting a breast US scan to identify its keyframes necessitates both a comprehensive understanding of normal and abnormal breast tissue to detect the lesion and accurate recognition of key sonographic features more specific to malignancy. To embed this clinical knowledge within the frame selection process while avoiding distracting the main task, we leverage an RL-based framework with a customized reward mechanism. Fig. 2 illustrates the whole pipeline. In this section, we first introduce the overall keyframe extraction framework for breast US video (denoted as KE-BUV), which learns from three sources of supervision. Then, we propose the detection-based nodule filtering (DNF) module that can filter out irrelevant content and tackle videos with unfixed lengths. Finally, we explain the attribute classification network (ACN) which is equipped with a novel group-aware cost function to handle the class imbalance issue.

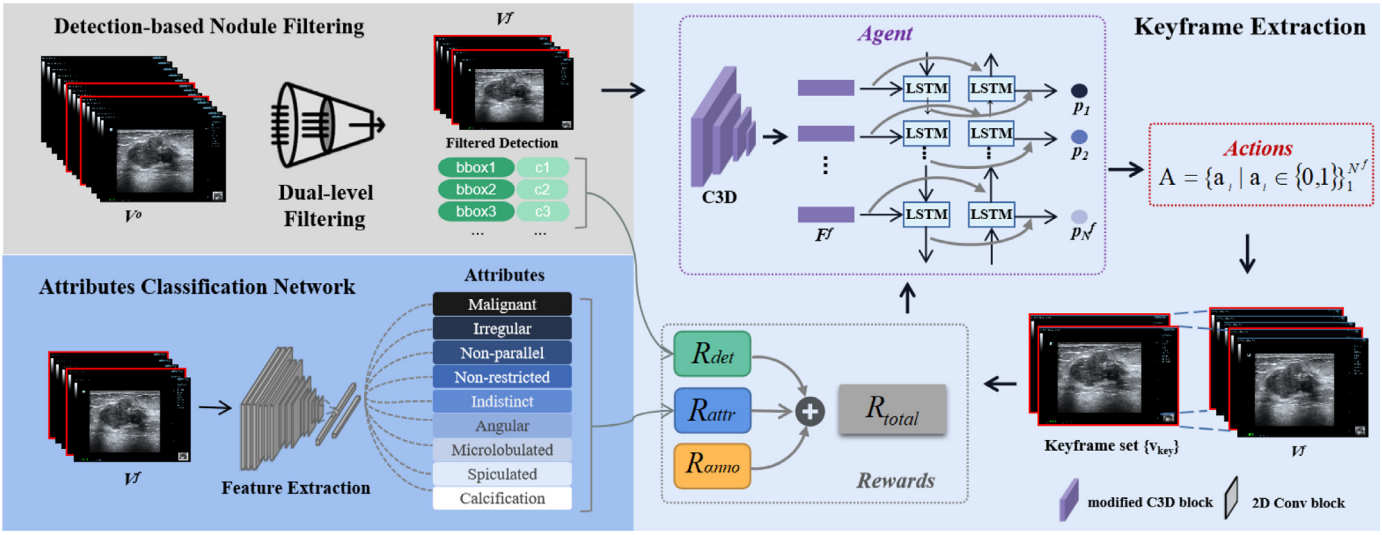

Fig. 2. Schematic of the proposed KE-BUV framework. A raw video V^o is first processed by the DNF module to filter out irrelevant content, yielding V^f and providing R_{det} reward. V_f is then passed to the agent to make suitable action $A = \{a_i | a_i \in 0, 1\}_1^{N^f}$ for each frame. The selected frames constitute the final keyframe set $\{v_{key}\}$. The agent is trained with reward signal that jointly considers expert annotation R_{anno}, nodule presence R_{det}, and malignancy indicator R_{attr}. Keyframes are highlighted in red, the purple cube represents the modified C3D block, while the grey cube indicates the 2D Conv block. Best viewed in color. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.1. The overall framework

The design of the KE-BUV framework is inspired by observing the sonographers' behavior during scanning. In specific, a sonographer first manipulates the probe to scan the breast to detect the lesion. Then, they adjust their focus on the nodular region to examine its size and location and recognize possible malignant indicators. Finally, the most representative frames are extracted to make the final diagnosis. Note that this procedure involves three intrinsically different tasks, i.e., detection, classification, and frame extraction and might require different features to solve. To better coordinate the learning between our primary task (i.e., frame extraction) and the two auxiliary ones, we integrate them through an RL-based model. The whole pipeline is shown in Fig. 2.

A raw breast US video can be denoted as $V^o = \{v^o\}_1^{N^o}$, with a size of $H \times W \times C \times N^o$, each represents height, width, color channel, and video length, respectively. V^o is first processed by the DNF module to eliminate irrelevant content and produce a filtered video $V^f = \{v^f\}_1^{N^f}, N^f \leq N^o$ with the size of $H \times W \times C \times N^f$. More details of the DNF module are explained in Section 3.2.

To generate a keyframe set with flexible size, we approach frame extraction as a sequential decision-making process. Given the common setting of RL: an agent in its current state $\boldsymbol{S}$, interacts with the environments $\mathcal{E}$ by making suitable action $a \in \mathcal{A}$ to maximize the expectation of reward R. Formally, we define the following elements:

Agent: The agent is a sampler that takes V^f as input and samples sequentially from a Bernoulli distribution to generate suitable action for every frame. It first maps the V^f to F^f using a modified C3D model[1] to extract high-level features. Then, F^f is further processed by a simple Bi-LSTM model (i.e., one Bi-LSTM layer with a hidden size of 256, one fully-connected (FC) layer, and a final sigmoid layer) to output probability p_i for frame i. p_i controls the likelihood of whether action '1' is selected for frame i. Note that existing works (e.g.(Zhou et al., 2018; Liu, Meng, Vlontzos, Tan, Rueckert, Kainz)) usually train the feature extractor and the likelihood prediction model separately, which may lead to disjunction in feature space. Instead, we train the two end-to-end to learn task-specific features.

Action: Action is defined as the set of frame-wise selection operation $A = \{a_i | a_i \in \{0, 1\}\}_1^{N^f}$, where '1' indicates selecting the current frame at time i, '0' indicates discarding.

State: State $S = \{v_{key}\}_1^{N^k}$ corresponds to the keyframe set that is selected by the agent. N^k denotes the number of frames that have been selected.

Reward: Based on S, the reward R evaluates the quality of the generated keyframe set and informs the agent what policy it should adopt to select the appropriate action. The proposed reward mechanism for KE-BUV framework jointly considers expert annotation R_{anno}, nodule presence R_{det}, and malignancy indicator R_{attr}.

R_{anno} reward the agent if the index of the predicted keyframe id_{key} is close to that of the manually annotated one id_{GT}. It is defined by:

$$R_{anno} = \frac{\sum_{key}^{N^k} Sim(id_{key}, id_{GT})}{N^k + \epsilon} \tag{1}$$

$$Sim(id_{key}, id_{GT}) = \begin{cases} \dfrac{\overline{id}_{key}^2}{\overline{id}_{GT}^2 + \epsilon} & id_{key} \leq id_{GT} \\ \dfrac{(\overline{id}_{key} - 1)^2}{(\overline{id}_{GT} - 1)^2 + \epsilon} & id_{key} > id_{GT} \end{cases}, \tag{2}$$

where $\overline{id}_{GT} = \frac{id_{GT}}{N^f}, \overline{id}_{key} = \frac{id_{key}}{N^f}$ are normalized by the video length. ϵ is a small number to avoid division by 0. The value of $Sim(id_{key}, id_{GT})$ ranges from [0,1], and reaches the maximum when $id_{key} = id_{GT}$. The quadratic function is adopted to amplify the performance increment. The final R_{anno} is the average of reward obtain by all elements in S. Note that the design of this reward considers the fact that only one id_{GT} annotation is available for each video in our dataset, while it can be easily extended to other scenarios.

R_{det} reward the agent if the lesion detected in frame v_{key} (denoted as $bbox_{key}$) is consistent with that detected in the ground truth frame v_{GT} (denoted as $bbox_{GT}$), or can be detected with a

[1] The temporal dimension of the C3D model is not down-sampled by using a pooling kernel of [2,2,1].

high confidence score. It is defined as:

$$R_{det} = \frac{\sum_{key}^{N^k} IoU(bbox_{key}, bbox_{GT}) + c_{key}}{N^k + \epsilon} \tag{3}$$

$IoU(\cdot)$ measures the intersection over union between the two bounding boxes, while c_{key} is the detection confidence score in frame v_{key}[2] The values of both the two elements range from [0,1]. R_{det} is calculated by averaging this reward across the whole keyframe set.

R_{attr} measures how many malignancy indicators can be recognized in the selected v_{key} for the malignant lesions, while suppressing this value for benign cases. The v_{key} is first fed to the attribute classification network to recognize whether the malignant indicators are present in this frame: $attr_{key} = ACN(v_{key})$ (details explained in Section 3.3). $attr_{key}$ is a 1×9 vector, whose elements corresponds to the probability of 9 different malignant indicators. Given the class of lesion $y \in 0, 1$ (i.e., '1': malignant, '0': benign), R_{attr} is defined as:

$$R_{attr} = \frac{\sum_{key}^{N^k} y * n_{key}^{MI} + (1-y) * (9 - n_{key}^{MI})}{N^k + \epsilon} \tag{4}$$

$$n_{key}^{MI} = sum(th(attr_{key}, 0.5)) \tag{5}$$

$th(\cdot, 0.5)$ thresholds the value with 0.5. n_{key}^{MI} calculates the total number of malignancy indicators in frame v_{key}.

The total reward is then defined as:

$$R_{total} = R_{anno} + R_{det} + R_{attr}. \tag{6}$$

Optimization process: The agent is trained to learn the policy function $\pi_\theta(a|S)$ w.r.t parameters θ by maximizing the total reward R_{total}. Following (Williams, 1992; Zhou et al., 2018), the derivative of the objective function $J(\theta)$ can be approximated as:

$$\nabla_\theta J(\theta) \approx \frac{1}{E} \sum_{e=1}^{E} \sum_{i=1}^{N^f} (R_e - b) \nabla_\theta \log \pi_\theta(a_i \mid h_i) \tag{7}$$

where e denotes the eth episode, E denotes the total number of training episodes. a_i, h_i represent the action and the hidden state at time i. b is the moving average of previous rewards to accelerate the training. $l2$ regularization is also applied to constraint both the model weights and the size of the keyframe set (Mahasseni et al., 2017). With a learning rate of α, θ can be updated as:

$$\theta_{i+1} = \theta_i + \alpha \nabla_\theta J(\theta) \tag{8}$$

Through iterative training, the agent learns to maximize rewards by approximating its prediction to the manual annotation, as well as selecting frames with clear visualization of the lesion and the diagnostic-related attributes. During the test, the proposed KE-BUV framework process an unseen video to not only yield a keyframe set but also provide a visual interpretation of the decision-making process by automatically labeling each keyframe with the detected lesion and the presence of malignancy indicators (e.g. example shown in Fig. 3). Comparison experiments show that it outperformed other state-of-the-art methods and its prediction aligned well with manual annotation ($dis_{best} = 1.97 \pm 0.10$, $dis_{ave} = 8.69 \pm 0.16$, $MD\text{-}3 = 83.22 \pm 0.46$, $MD\text{-}10 = 97.44 \pm 0.85$). Moreover, the reader study validates that the produced keyframe set can help to improve the diagnosis accuracy (see Section 5.6).

[2] The generation of $bbox_{key}$ and c_{key} is explained in Section 3.2.

3.2. Detection-based nodule filtering module

Compared with a single still image, a breast US video contains copious information about both the lesion and the surrounding tissues. Analyzing the full video to recognize the lesion is arduous, however, as the composition of breast tissue varies across individuals by age, menopausal status, ethnicity, and weight (Boyd et al., 2010). Another leading issue is the existence of redundant information, as a lesion might only be visible within a small portion of the video. The length of the videos can also be drastically different (ranging from 50 to more than 400 frames). To address these, we propose a detection-based nodule filtering (DNF) module. It can eliminate irrelevant information and helps to tackle videos of different lengths (see Fig. 4).

Specifically, frame $v_i^o \in V^o$ is first processed by a pre-trained Yolo model (Redmon and Farhadi, 2018) (details explained in Section 4.3) to detect the lesion in 2D frames, producing a 5 dimensional vector $bbox_i^o = \{x_i, y_i, h_i, w_i, c_i\}$, while x_i, y_i corresponds to the location of the detected lesion, h_i, w_i represents the size, c_i is the confidence score. Due to the various challenges in detecting breast lesions, this process may result in oscillating bounding box predictions across the whole video (the blue line in Fig. 5). We point out that is also possible to replace this model with 3D or video-based ones while collecting annotations to train such models is tedious and expensive. Therefore, the 2D detection model is used for its efficiency and being less annotation-hungry.

To handle the oscillating predictions, we use dual-level filtering to attenuate the noise. First, window-level filtering is performed following:

$$\bar{c}_i^o = \begin{cases} \frac{\sum_0^i c_i}{i} & i < l \\ \frac{\sum_{i-l}^{i+l} c_i}{2l} & l \le i < N^o - l \\ \frac{\sum_{i-l}^{N^o} c_i}{N^o - i + l} & N^o - l \le i \end{cases} \tag{9}$$

, where $2l$ is the window width, and $\bar{c}_i^o$ is the smoothed confidence score for frame i. The key insight here is to replace the original c_i^o with the average confidence score across the current window. It can produce smoothed prediction and is robust to random noise (e.g., the red curve in Fig. 5). Next, the video-level filtering is conducted via first sorting the $\{\bar{c}^o\}_1^{N^o}$ in the descending order. Then the top K frames with the highest value are selected, and the average of their index is selected as the final window anchor id_c. Formally:

$$\begin{aligned} id_c &= \frac{\sum\{id\}_1^{2l'}}{2l'} \\ \{id\}_1^K &= argsort(\{\bar{c}^o\}_1^{N^o}, 2l'), \ s.t. \ \bar{c}_{id_1}^o \ge \bar{c}_{id_2}^o \ge \dots \bar{c}_{id_{2l'}}^o. \end{aligned} \tag{10}$$

$argsort(\cdot)$ function returns the indices of the top $2l'$ elements in a set. $2l'$ is the hyper-parameter that controls the video-level smoothness. The filtering process can exploit the lesion detection information across the whole video and further enhance the model robustness. Finally, the DNF module returns a filtered video with a fixed length of $2l'$:

$$V^f = \begin{cases} \{v^o\}_1^{2l'} & id_c < l' \\ \{v^o\}_{id_c - l'}^{id_c + l'} & l \le id_c \le N^o - l' \\ \{v^o\}_{N^o - 2l'}^{N^o} & id_c < N^o - l' \end{cases} \tag{11}$$

This piece-wise function is defined to avoid exceeding the time limits. Note also that $2l' = N^f$, and any raw video V^o with a frame number of $N^o < 2l'$ is up-sampled uniformly to $2l'$ before further processing. In this work, we set $l = 25, l' = 50$ to have a sufficiently large searching window for subsequent keyframe extraction.

Empowered by the DNF module (see Fig. 4), the proposed framework can accept inputs of flexible sizes. It also avoids exces-

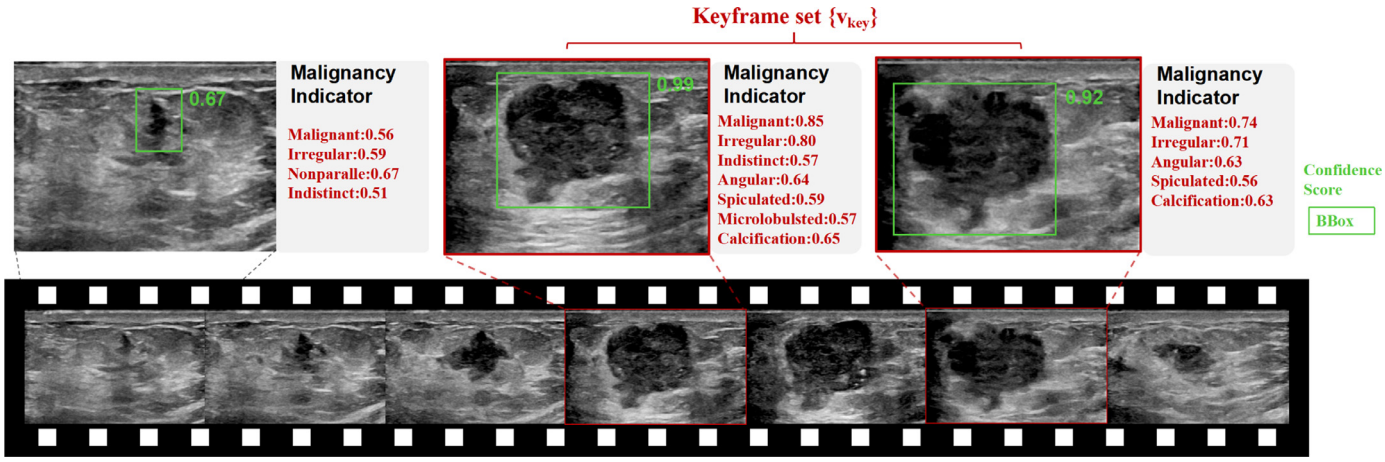

Fig. 3. Test output of the proposed framework. Corresponding frames from a breast US video are plotted sequentially at the bottom. Detected keyframes are highlighted in red. The keyframes are accompanied by the detected lesion bounding box (green rectangle) and the predicted likelihood of different malignancy indicators (listed in red). This information can then assist the user to perform the further examination and making the diagnosis. The frame plotted in the upper-left corner demonstrates the case where a frame is not selected as it results in a small gain in R_{det} and R_{attr}. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

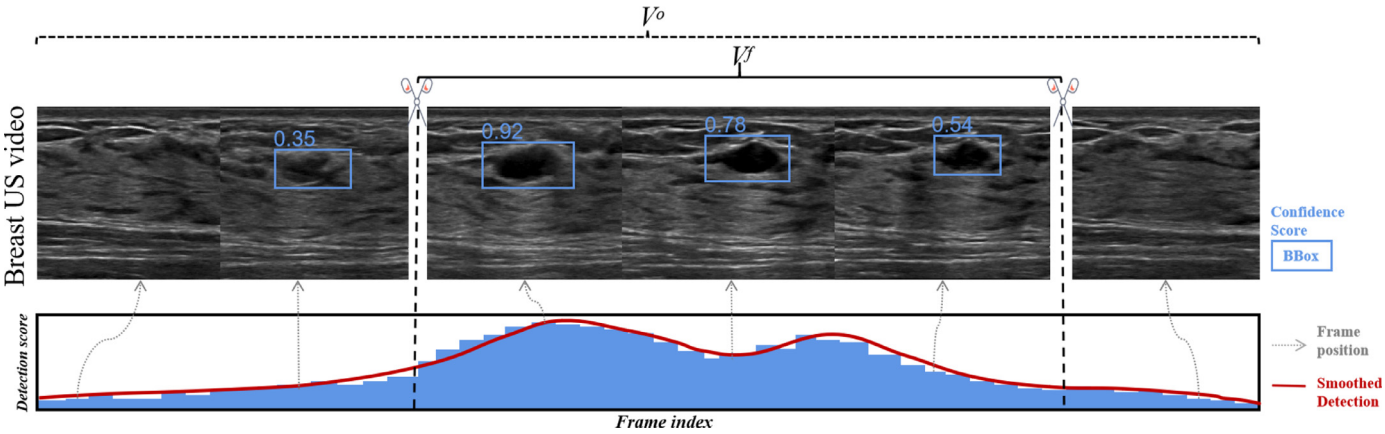

Fig. 4. Illustration of the DNF module. Given a raw video V^o with an unknown length, each of its frames is processed by the lesion detection model to yield a bounding box (blue rectangle) and a confidence score. The DNF then applies dual-level filtering to smooth the predictions (details refer to the text). It finally eliminates irrelevant information and produces a filtered video V^f with a fixed length. Note that the histogram shown in this figure is simplified for demonstration purposes, while real-world examples of this plot can be referred to in Fig. 5. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sive down-sampling of videos and auto-focus the keyframe searching to the most relevant region. V^f is then forwarded to subsequent modules for further processing. Similarly, $bbox_j^f$ corresponds to the j element $v_j^f \in V^f$ is also extracted to calculate the R_{det} (see Eq. (3)).

3.3. Attribute classification network with group-aware focal loss

Another key factor to consider when selecting keyframes for breast US videos is how many malignancy indicators can be visualized in these frames. To transfer this prior knowledge to the keyframe extraction task, we train an attribute classification network (ACN) to automatically identify the malignancy indicators and use its prediction to guide the agent in action selection. Based on the finding of several clinical studies (JM et al., 2009; Paulinelli et al., 2005; Kim et al., 2010; Chao et al., 1999; Heinig et al., 2008), we selected 8 sonographic attributes (i.e., irregular shape, non-parallel orientation, non-restricted boundary, indistinct margin, angular margin, micro-lobulated mass, spiculated mass, and micro-calcification) as they indicate malignancy with relatively high odds ratio. The proposed ACN then takes a single US frame as input and predicts whether the lesion exhibits these attributes and whether it is malignant (9 outputs in total).

However, solving this multi-label classification problem is non-trivial as it faces severe class imbalance issues (e.g. a class ratio of 87% vs. 13%). This could lead to gradient dominance in the majority class and yield biased prediction (Johnson and Khoshgoftaar, 2019). To mitigate this problem, we propose the Group-aware focal loss (GAFL), which is defined as:

$$\mathcal{L}_{CFL} = -[y(1-p_+)^{\gamma_+} log(p_+) + (1-y)p_-^{\gamma_-} log(1-p_-)], \tag{12}$$

$$p_+ = min(p, th_+),\ p_- = max(p, th_-), \tag{13}$$

where y represents the ground truth label for each class, and p is the predicted score. γ_+, γ_- are hyper-parameters that adjust weights for the positive (minority) class and the negative (majority) class, respectively. th_+, th_- are thresholds to eliminate the effects of easy samples. The design of GAFL follows two observations:

(1) The distribution of easy vs. hard examples within the majority group is incompatible with that of the minority group. For

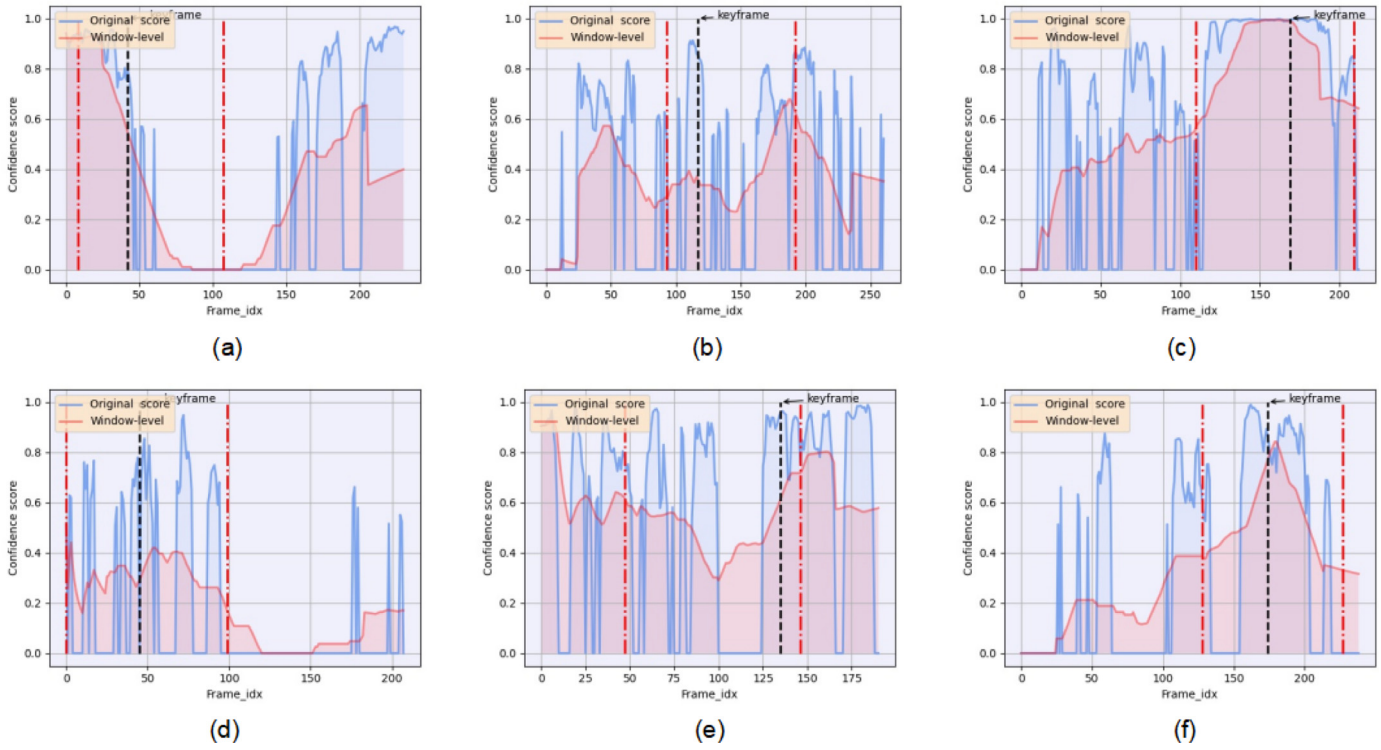

Fig. 5. Examples before/after the dual-level filtering process. The blue line indicates the confidence score of different frames within each video. The red curve represents the window-level filtered results. The red dashed line indicates the position of V^f obtained after dual-level filtering. The black dashed line represents the position of the manually selected keyframe. It shows that the dual-level filtering process can smooth the oscillating predictions and help the model to focus on the critical region for keyframe extraction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

example, easy examples might dominate the majority group, while the minority group witnesses a more equal distribution. Therefore, the focusing parameter γ (initially proposed in Lin et al. (2017)) should be further segregated for different groups. In specific, γ_- is set with a higher value to down-weight easy examples present in the majority group.

(2) Due to its sheer amount, negative examples dominate the total gradients during network parameter updating and cause distraction in learning the more interested, challenging examples. To further counteract this, we adapt simple min/max operation to diminish gradient back-propagation for these simple negative examples. In other words, $\frac{\partial \mathcal{L}}{\partial p} = 0$, given $p < th_-, y = 0$, and vice versa.

Equipped with GAFL, the ACN is first trained on the labeled 2D dataset and is subsequently applied to different frames in $\{v_{key}\}$ to calculate the R_{attr} to guide the keyframe extraction (see Eq. (4)). Note that as these attributes can independently present, 9 GAFL are assigned to 9 different attributes and they are trained simultaneously (see Fig. 2). Note also that attribute classification has been investigated by other studies (e.g.,Liu et al., 2020b; Zhang et al., 2020; Hernández-López and Gómez-Flores, 2020). In this work, we opt for a simple yet classical model to fulfill this auxiliary task while this model could be replaced in other applications as the proposed GAFL is general.

Relation with Ridnik et al.(2021) A recent work of Ridnik et al. (2021) also introduced a loss function that decouples the γ as the proposed *GAFL*. However, they further proposed to subtract $m = 0.2$ from p to shift the whole probability distribution $p_m = \max(p - m, 0)$ (given $y = 0$). While this shifting can also diminish the effects of easy negative examples, it also hampers the learning of hard examples and leads to false-positive predictions. On the contrary, the proposed GAFL does not have this limitation. Furthermore, GAFL applies a similar strategy to the minority group as some of the attributes witnessed a more even distribution (see Section 4.1). Later comparison experiment shows that the GAFL outperforms (Ridnik et al., 2021) and other competing methods in recognizing highly imbalanced lesion attributes.

4. Experiments and materials

4.1. Dataset

The video dataset was obtained from 653 patients at the following hospitals: Sun Yat-sen Memorial Hospital; Gansu Provincial Cancer Hospital; The Second Affiliated Hospital of Chongqing Medical University; Central People's Hospital of Zhanjiang; Boai Hospital of Zhongshan City; Affiliated Hospital of Guangdong Medical University; The First Affiliated Hospital of Nanchang University; West China Hospital of Sichuan University. The study was approved by the local Institutional Review Board. In total, 2606 videos of breast lesions were acquired using different US machines produced by Philips, Mindray, Toshiba, etc. All videos were resized to have a height of 400 and a width of 600. To simplify the annotation process, one keyframe was annotated for each video by experienced sonographers. A bounding box of the lesion and its corresponding clinical attributes (i.e., irregular shape, non-parallel orientation, non-restricted boundary, indistinct margin, angular margin, microlobulated margin, spiculated margin, and calcification) were also labeled on this frame. Biopsies were carried out for all the patients and the results are used as the ground truth label for lesion malignancy. The distribution of each attribute is listed in Table 1. It can be seen that most attributes witness severe class imbalance (e.g. spiculated vs. non-spiculated: 7.56:1). The whole dataset is partitioned at the patient level to a 7:1:2 split for training validation, and testing. The annotated keyframes were extracted and served as the corresponding 2D training dataset for the DNF module and the ACN.

Table 1
Distribution of different attributes of the breast lesions.

Type	Attributes	Ratio
Cancer	Benign/Malignant	2.74:1
Shape	Regular/Irregular	1.29:1
Orientation	Parallel/ Non-parallel	4.77:1
Margin	Restricted/Non-restricted	1.26:1
	Clear/Indistinct	3.28:1
	Non-/Angular	3.85:1
	Non-/Spiculated	7.56:1
	Non-/Micro-lobulated	3.81:1
Calcification	Non/Micro-calcification	3.36:1

4.2. Experiments

Comparison experiments To demonstrate the effectiveness of the proposed framework, we compared the proposed model with both classical and the state-of-the-art approaches (i.e., VSLSTM (Zhang et al., 2016), GANdpp (Mahasseni et al., 2017), FCSN (Rochan et al., 2018), USPD (Chen et al., 2017), VSANet (Fajtl et al., 2018), DeepRL (Zhou et al., 2018),UVS (Liu et al., 2020a), AFUSPR (Pu et al., 2021)). Among them, USPD, UVS, and AFUSPR were originally proposed to process US videos, DeepRL and UVS also utilized an RL-based framework. We use the same $l2$ regularization to constraint the size of the keyframe set (Mahasseni et al., 2017) to be less than 6 for all RL-based methods. For all other comparing methods, we take 6 keyframes with the highest prediction score to form the corresponding keyframe set. Note that the evaluation metrics used in this study (see Section 4.4) used the closest distance between the predicted keyframe set and the GT in quantitative evaluation. As a result, the non RL-based methods has additional advantages in these metrics as they have a larger keyframe set size ($N_{key} = 6$) than that of the RL ones ($N_{key} \leq 6$). We perform a suitable hyper-parameter selection for all competing methods. More implementation details of the proposed framework could be found in Section 4.3. All experiments are conducted under 5-fold cross-validation and data separation was conducted at the patient level to avoid data contamination. Quantitative results are shown in Table 2.

Ablation study One of the key contributions of this work is the specially designed reward mechanism. To investigate the effect of each of its components, we carry out ablation study by training the agent using R_{anno} only, $R_{anno} + R_{det}$, $R_{anno} + R_{attr}$, and the full framework ($R_{anno} + R_{det} + R_{attr}$). Other hyper-parameters and network architecture are kept the same for a fair comparison. Results are reported in Table 3. All experiments are conducted under 5-fold cross-validation.

Agent backbone selection

Another factor that could affect the performance is the model architecture of the agent. To examine this, we implement several popular deep learning models in video analysis, i.e., R2D, R(2+1)D, C3D, P3D, ResNet+Bi-LSTM, C3D+Bi-LSTM. Quantitative results are reported in Table 4. The 3D models (e.g. C3D, R2D, R(2+1)D, etc.) are modified to take the input of $400 \times 600 \times 100$, and output a 1×100 vector for the subsequent sequential sampling. Similarly, the ResNet+BiLSTM model first processes each frame within a video as 2D inputs, and concatenates all predictions to feed into a BiLSTM model (one BiLSTM layer with a hidden size of 256 and an FC layer). It then outputs the final prediction with the size of 1×100 vector. Details of the C3D+BiLSTM model can be referred to in Section 3.1.

Impact of dual-level filtering The proposed DNF module uses dual-level filtering to smooth oscillating bounding box prediction and remove redundant information. To reveal the impact of its component, we implement the DNF module without any filtering (vanilla version); with only the window-level filtering; and with dual-level filtering (ours). In specific, the vanilla DNF picks the frame with the highest detection confidence score as the center and crops a clip with the length of $2l'$ to avoid perturbing the temporal dimension. The window-level DNF uses a similar strategy while it utilizes a smooth confidence score (explained in Section 3.2). Note that the three modules use the same detection model to produce frame-level prediction results during the test, while the filtering mechanism is different. More implementation details of the DNF module are listed in Section 4.3. To better examine the DNF, we also compare it to uniform sampling where all videos are up-sampled or down-sampled to the same temporal size. As the lesions and the corresponding keyframes could appear in varying locations within US videos, we calculate the percentage of filtered videos that contains (or not) the manually annotated frame (i.e., 'contains GT' and 'without GT' in Table 5). dis_{best} is also used to calculate the minimal distance between the filtered video and the GT frame (i.e. $dis_{best} = 0$ indicates that the GT is contained in the output).

Efficacy of Group-aware Focal Loss Another key component that needs to be evaluated is the efficacy of the proposed GAFL loss. We first carry out comparison experiments using identical backbones and trained with different loss functions (i.e., cross-entropy (CE) loss, the Focal loss (Lin et al., 2017), the ASL loss Ridnik et al. (2021), and the proposed GAFL loss). The same feature extraction backbone and training parameters are used. We performed a suitable hyper-parameter selection for all competing methods. More detail regarding the implementation is explained in Section 4.3. Results are reported in Table 6.

Meanwhile, it is also worthwhile to reveal the influence of the key hyper-parameters of the GAFL loss. We first implement the GAFL loss with different values of $\gamma_{-,+}$, while keeping $th_{+,-} = 0$. Note that we set $\gamma_- \geq \gamma_+$ to attenuated the effect of the majority negative class (e.g. $\gamma_- = 1, 2, 3, 4$, $\gamma_+ = 1$). To verify this assumption, we also test the performance by reversing the value: $\gamma_+ = 3, \gamma_- = 1$. Then with the selected $\gamma_{+,-}$, different pairs of $th_{+,-}$ are also tested. Experimental results are reported in Table 7.

Reader Study To further validate the efficacy of the proposed framework, we conducted additional reader studies to test whether doctors are qualitatively satisfied with the selected keyframes and whether the automated keyframe selection would benefit the diagnosis. In specific, 3 junior clinicians (2 years of experience) and 3 clinical experts ($\geq$ 6 years of experience) are invited to participate in two tasks:

(1) Quality rating; 50 breast US videos are randomly selected and fed to the KE-BUV framework to automatically generate keyframe sets. Participants are then shown each of these videos and the corresponding keyframe set. They then rate the quality of the keyframe set to 'Excellent', 'Fair', and 'Poor'. Results are shown in Table 8.

(2) Cancer diagnosis. It is interesting to investigate whether the automated generated keyframe sets will benefit the subsequent diagnosis. To examine this, participants from different groups are shown with different inputs to perform the diagnosis (classify the lesion into benign and malignant). In specific, we use three types of inputs: a. static 2D US image of a lesion (manually selected keyframe of each lesion). b. automated generated keyframe set. c.the whole US video of a breast lesion. We randomly sampled 50 cases from the test set for this experiment. Data are re-shuffled and sent to participants at one-month intervals. Results are shown in Table 9.

4.3. Implementation details

All experiments were implemented in PyTorch with a GeForce RTX 3090 GPU. The same dataset partition was maintained to train

the DNF, ACN, and the full KE-BUV framework during the cross-validation experiments to avoid data contamination. For comparison methods that only accept videos with a fixed length, we uniformly down-sample or up-sample videos to 100 frames (i.e., uniform-sampling) for a fair evaluation. *KE-BUV framework* The agent is trained from scratch with an Adam optimizer using a learning rate of 1e-4 and a weight decay of 1e-5. Other hyper-parameters were set as $E = 5, N^f = 100, \epsilon = 1e-4$ (see Eqs. (1), (3), (4) and (7).

DNF module As labeling the whole video is tedious and time-consuming, we trained a Yolo model to automatically detect breast lesions using the annotated 2D dataset (selected keyframes extracted from the videos). An Adam optimizer is used with a learning rate of 1e-4 and a weight decay of 5e-4. Model weights were initialized using the COCO pre-trained weights. Random flipping, cropping, and brightness modification were applied as data augmentation. Note that other detection models could also be applied, while we opted for Yolo-v3 for its efficiency and robustness. The trained 2D detection model is then applied to different frames to analyze the whole video. Other hyper-parameters were set as: $l = 25, l' = 50$ in this experiment. These numbers can be easily customized to analyze different data.

ACN The ACN was trained using the proposed GAFL as the objective. A ResNet-18 model is trained with an Adam optimizer using a learning rate of 1e-3 with a weight decay of 1e-4. Model weights were initialized using the ImageNet weights. Random flipping, rotation, scaling, and color jittering were applied as data augmentation. Similarly, the annotated 2D dataset was used to train the ACN. The trained model is then fixed and is applied iteratively to analyze every frame across a video in the KE-BUV framework. Hyper-parameters were set as: $\gamma_+ = 1, \gamma_- = 3, th_+ = 0.15, th_- = 0.2$ for GAFL (more details explained in Section 5.5), $\gamma = 1.5$ for Focal loss, and $m = 0.2, \gamma_+ = 0, \gamma_- = 2,$ for ASL loss. Details of the hyper-parameter tuning process for the Focal and ASL loss are omitted as it is not the main focus of this paper.

4.4. Evaluation metrics

As explained earlier, only one representative keyframe is labeled for each video to simplify the annotation process. Due to this extreme sparse annotation, the model performance cannot be evaluated using common metrics such as precision or F1-score adopted in (Liu et al., 2020a; Chen et al., 2017). An ideal keyframe set should include elements that are close to the GT, while diversity may also be encouraged to fully represent the lesion. Therefore, we define the following metrics:

$$dis_{best} = \min_{key \in \{v_{key}\}_1^{N^k}} |id_{GT} - id_{key}|, \tag{14}$$

$$dis_{ave} = \frac{\sum_{key \in \{v_{key}\}_1^{N^k}} |id_{GT} - id_{key}|}{N^k}, \tag{15}$$

where $|id_{GT} - id_{key}|$ calculates the distances between the ground truth frame and the key_{th} keyframe, $\{v_{key}\}_1^{N^k}$ is the keyframe set. dis_{best} measures the minimal distance between the predicted keyframe set to the manually annotated one, while dis_{ave} reports the averaged distance between predicted keyframes and the GT frame.

As adjacent frames often exhibit similar information, we further loose the constraint by calculating the percentage of data whose minimal distance between $\{v_{key}\}$ and the GT frame is smaller than a given threshold δ. Formally,

$$MD\text{-}\delta = \frac{sgn(\delta - dis_{best})}{N^{test}} \times 100\%. \tag{16}$$

$sgn(\cdot)$ denotes the sign function. δ is set as 3,5,10, respectively (corresponds to $MD\text{-}3, MD\text{-}5, MD\text{-}10$). N^{test} represents the total number of test data. We also calculate $1 - MD\text{-}10$ to evaluate cases where the prediction has a large deviation from the ground truth (the smallest dis_{best} is larger than 10).

For the attribute classification experiment and the reader study, we use the AUC (area under the ROC curve), accuracy, sensitivity, precision, specificity, F1-score, G-Mean to evaluate the classification performance.

5. Results and discussion

5.1. Comparison experiments

The results of the comparison experiments are exhibited in Table 2. The proposed KE-BUV framework scored superiorly in all metrics ($dis_{best} = 1.97 \pm 0.10$, $dis_{ave} = 8.96 \pm 0.16$, $MD\text{-}3 = 83.22 \pm 0.46\%$, $MD\text{-}10 = 97.44 \pm 0.85\%$). This shows that our approach can yield a keyframe that is aligned with the manual annotation, and the average best distance is less than 2. It might be explained that it is the first approach bespoke for the breast US data which incorporates relevant clinical knowledge into its model design. The classical VSLSTM obtained the worst dis_{best} and $MD\text{-}3$, suggesting that detecting keyframes of breast lesions is a challenging task and a simple combination of CNN and LSTM cannot handle this problem. It can be seen that GANdpp obtained higher performance than VSLSTM. This may result from the additional constraint of GANdpp which ensures similarity between the video feature and the keyframe features. Interestingly, FCSN fails to achieve similar performance. We conjecture that this could cause by the sparse annotation available for this work which yields inferior performance when formulating as a segmentation task. VASNet was also able to score higher performance than VLSTM. It indicates that information across the temporal dimension might be critical in finding the keyframes. Meanwhile, USPD exhibits better accuracy than VSLSTM and GANdpp. This might verify the importance of end-to-end training in the keyframe extraction task which helps the backbone to learn task-specific features. Similarly, AFUSPR also used end-to-end training and scored even better accuracy due to its additional optical flow stream. Note that the computation of optical flow can be time-consuming and might add additional computation burdens. Interestingly, the DeepRL model achieved good performance while it did not utilize the position of the GT keyframe. Meanwhile, UVS exceeded DeepRL as it introduces additional rewards indicating the presence of relevant structures. However, our approach could directly utilize manual annotation through R_{anno}, and can further incorporate diagnostic related features through R_{attr}, which leads to preferable performance.

Visualization results are displayed in Fig. 6. Fig. 6 (a) shows the original breast US video. The manually selected GT keyframe is plotted in red. The predicted keyframes by different comparing methods are listed in sub-figures (b)–(j), respectively, and their corresponding spatial locations are highlighted in different colors. The VSLSTM performed the worst in this case and selected the first and the last few frames as keyframes. The GANdpp model, on the other hand, selected the last few frames. Both these models failed to find keyframes that contain the lesion. We conjecture that this may cause by the varying appearance of lesions that are difficult to recognize. The FCSN and the VASNet method acted differently from the GANdpp model and selected keyframes at large intervals. This may be explained by the fact that only sparse annotation was available during the training and these models have learned to overfit the training set by predicting frames at specific positions. The USPD model selected adjacent frames in the middle of the video. It can be seen that the prediction score for each frame is close to the other within that range. The lesion is partially visible in these frames but they do not reveal the key features of this lesion. The AFUSPR model also rated frames in the

Table 2
Results of comparison experiments. Each row corresponds to different methods. The definition of the evaluation metrics is explained in Section 4.4. The best results are highlighted in bold.

Method	dis_{best}	dis_{ave}	*MD*-3(%)	*MD*-5(%)	*MD*-10(%)	1-*MD*-10(%)
VSLSTM	7.24 ± 0.05	16.08 ± 0.05	34.95 ± 0.38	48.39 ± 0.50	74.79 ± 0.29	15.21 ± 0.29
GANdpp	4.64 ± 0.44	12.16 ± 0.84	54.32 ± 2.98	67.79 ± 3.56	88.12 ± 3.49	11.88 ± 3.49
FCSN	5.12 ± 0.08	14.82 ± 0.10	51.61 ± 0.43	62.81 ± 0.59	83.70 ± 1.00	16.30 ± 1.0
USPD	3.75 ± 0.20	10.01 ± 0.21	64.07 ± 1.54	75.47 ± 1.60	90.02 ± 0.98	9.98 ± 0.98
VSANet	3.41 ± 0.22	13.55 ± 0.16	66.34 ± 0.42	80.94 ± 1.50	96.20 ± 0.93	3.80 ± 0.93
DeepRL	3.57 ± 0.22	15.45 ± 0.42	62.26 ± 1.42	77.81 ± 2.22	94.03 ± 1.45	5.97 ± 1.45
UVS	3.04 ± 0.09	14.91 ± 0.20	68.33 ± 1.56	83.57 ± 1.35	96.08 ± 0.59	3.92 ± 0.59
AFUSPR	3.02 ± 0.12	9.91 ± 0.19	70.05 ± 1.32	79.48 ± 1.21	92.17 ± 0.45	7.83 ± 0.45
Ours	$\mathbf{1.97 \pm 0.10}$	$\mathbf{8.69 \pm 0.16}$	$\mathbf{83.22 \pm 0.46}$	$\mathbf{91.40 \pm 0.57}$	$\mathbf{97.44 \pm 0.85}$	$\mathbf{2.56 \pm 0.85}$

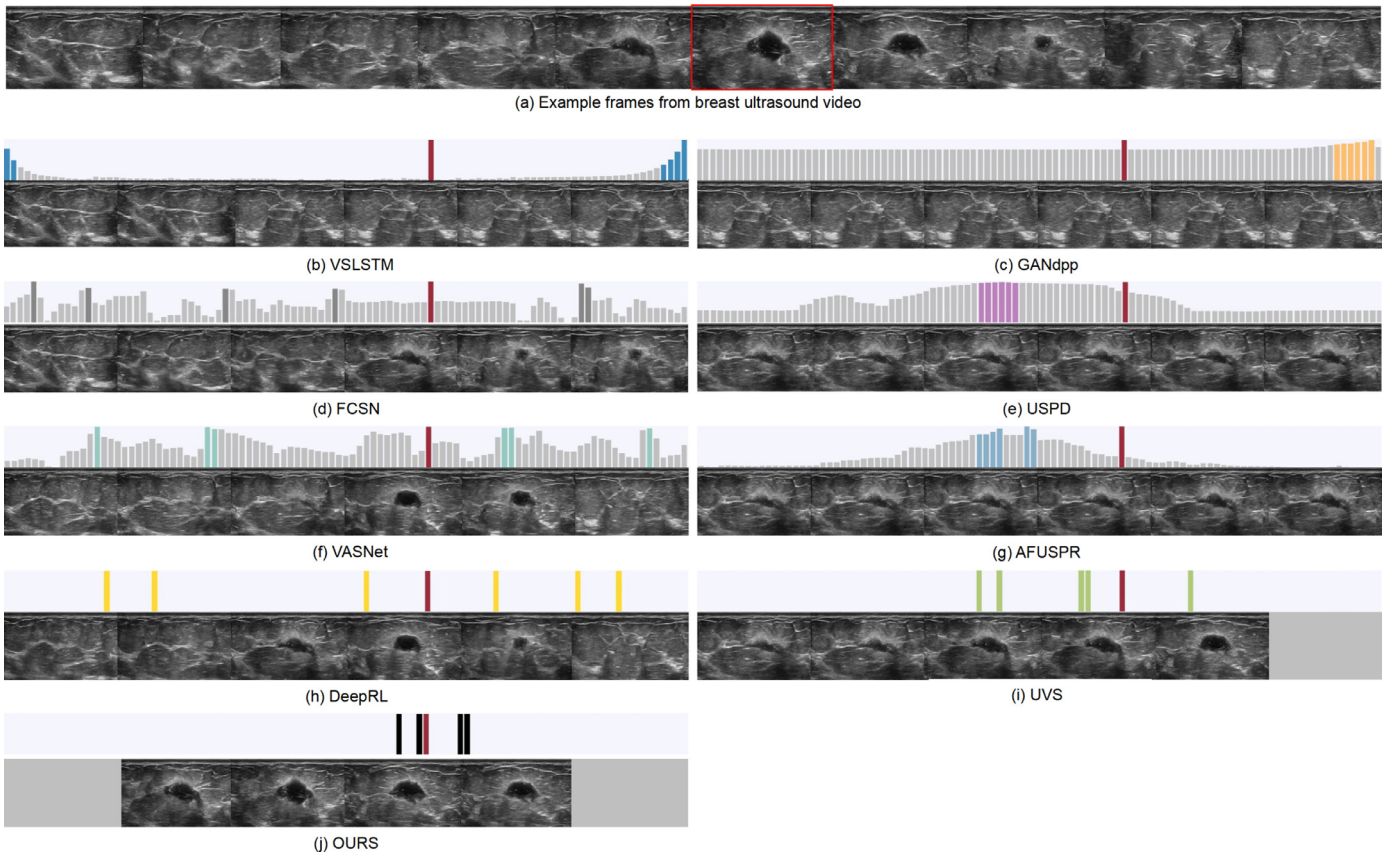

Fig. 6. Visual results of the comparison experiment. (a) shows the original breast US video. The manually selected GT keyframe is plotted in red. The predicted keyframes are listed in sub-figures (b)–(j) and their corresponding spatial locations are highlighted in different colors. We also plotted the prediction score for each frame in grey when applicable. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

middle higher than the others. It can be seen that it is more selective than the USPD model. Note that one difference between a common natural image video and a breast US video is that the position of the camera is often fixed during the data collection for the former. On the contrary, the US probe moves across the breast during the examination. As a result, the background (non-lesion region) and the foreground (the lesion) both change across a breast US video. Therefore, the additional optical flow input used in the AFUSPR model might focus on incorrect frames. The DeepRL, UVS, and the proposed framework were all able to select frames with clearer visualization of the lesion. The UVS performed better than the DeepRL as it also introduced additional guidance regarding the presence of the lesion. However, neither of them was able to utilize the annotation information directly and some of the keyframes they selected do not represent the whole lesion. The proposed framework, however, was able to utilize the sparse annotation and output a keyframe set of higher quality.

5.2. Ablation study

Table 3 exhibits the impact of each component used in the proposed reward function. The first row shows the result of training only using the R_{anno}. Note that its performance is already higher than some of the competing methods such as DeepRL, GANdpp, FCSN, etc. By introducing the R_{det} reward, the model performance can be boosted by a large margin as it provides critical information about the presence of the lesion. This helps to avoid distraction from redundant information or misleading surrounding tissues, and focus the keyframe extraction among highly relevant frames. Moreover, the full framework further incorporates the diagnostic-related attribute information through R_{attr} and can attain even better consistency with the experts in keyframe selection (e.g. row 3 in Table 3). This is in line with clinical practice for breast US examination, where the existence of malignancy indicators also influences the frame selection process. Interestingly, $R_{anno} + R_{det}$ scored

Table 3
Results of the variants of the proposed framework trained with the R_{anno} only, $R_{anno} + R_{det}$, $R_{anno} + R_{det}$, and the full framework ($R_{anno} + R_{det} + R_{attr}$), respectively.

Component	dis_{best}	dis_{ave}	*MD*-3(%)	*MD*-5(%)	*MD*-10(%)	1-*MD*-10(%)
R_{anno}	3.31 ± 0.15	13.21 ± 0.11	63.03 ± 1.40	80.00 ± 1.07	94.17 ± 1.01	5.83 ± 1.01
$R_{anno} + R_{det}$	2.25 ± 0.14	$\mathbf{8.44 \pm 0.17}$	80.49 ± 1.52	89.10 ± 0.53	95.59 ± 0.27	4.41 ± 0.27
$R_{anno} + R_{attr}$	3.06 ± 0.26	12.70 ± 0.88	66.02 ± 1.95	82.46 ± 1.88	95.22 ± 0.84	4.78 ± 0.84
Full framework	$\mathbf{1.97 \pm 0.10}$	8.69 ± 0.16	$\mathbf{83.22 \pm 0.46}$	$\mathbf{91.40 \pm 0.57}$	$\mathbf{97.44 \pm 0.85}$	$\mathbf{2.56 \pm 0.85}$

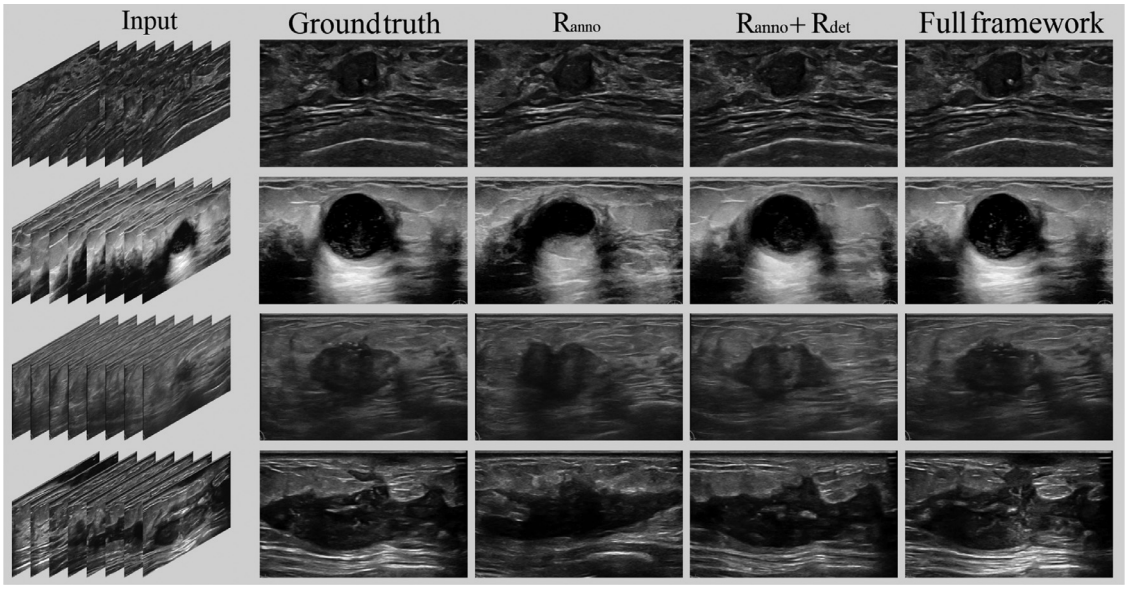

Fig. 7. The visual result of the ablation study. Different rows represent different cases. The second column displays the manually selected GT keyframes. To reduce the figure size, we only display the best predictions generated by different models (i.e. the closet to the GT in each predicted keyframe set). Figures are zoomed near the lesion region for easier visualization.

a smaller dis_{ave} than that of the full framework. This might cause by the introduction of the R_{attr} reward which encourages selecting frames with distinct diagnostic attributes. These frames also contain crucial information while they might be relatively far from the single GT frame. We also implement the same framework using $R_{anno} + R_{attr}$. It can be seen that it outperformed the vanilla R_{anno}-only version, while being inferior to that trained with $R_{anno} + R_{det}$. This might be explained that without the guidance of R_{det}, R_{attr} might be distracted by misleading surrounding tissue that does not belong to the lesion. Note that the full framework still produces smaller dis_{best}, and higher *MD*-3, 5, 10. This might demonstrate that the full framework can produce keyframe sets that are better aligned with the manual annotations while being more diverse at the same time.

Fig. 7 provides qualitative results of the ablation study. It can be seen that the automated selected keyframes are visually similar to the GT frame. Meanwhile, $R_{anno} + R_{det}$ produced predictions with a clearer view of the lesion than that of the R_{anno} (e.g., row 2 and 4). This might result from that the appearance of a lesion could change quickly due to its irregular shape or fast probe manipulation during the data collection. As a result, a frame that is not far away from the GT (i.e., has a high R_{anno}) may not provide produce a high R_{det} reward. This validates the necessity of including additional lesion presence information through R_{det}. Meanwhile, the predictions generated by the full framework and that of the $R_{anno} + R_{det}$ are overall similar (column 3 and 4). However, the former includes diagnostic attribute information and can perform even better in keyframe selection. For example, row 1 and 4 of Fig. 7 exhibits cases where the full framework can pick frames with the presence of calcification (bright spot within the lesion). It is in accord with that selected by the expert (column 2, Fig. 7). This subtle difference might be critical to further examination and influence the subsequent diagnosis. This suggests that all components of the proposed novel reward mechanism are beneficial and should be jointly considered during the keyframe selection.

5.3. Agent backbone selection

Table 4 demonstrates the impact of agent architecture on the model performance. It can be seen that most variants scored higher accuracy than other competing SOTA methods (shown in Table 2). This suggests that the proposed framework is robust and its overall design is well-suited for detecting keyframes in breast US. Interestingly, the R(2+1)D backbone obtained lager dis_{best} and $1 - MD$-10. We conjecture that this disentanglement of spatial and temporal convolution might have hampered the model performance. Also, note that it produces the smallest dis_{ave} among the six models. A brief examination of its results shows that this model tends to select adjacent frames. The C3D and P3D backbone performed superiorly to R2D, while is inferior to ResNet+BiLSTM. This might suggest that the keyframe selection task may benefit more from long-range temporal information captured by the BiLSTM. Among different variants, C3D+BiLSTM performed the best and is used in other comparison experiments. Its success might stem from a combination of 3D convolution and BiLSTM which enables joint learning across the spatial and temporal dimensions. We opt for this agent backbone for other experiments.

5.4. Impact of dual-level filtering

Table 5 exhibits the impact of dual-level and window-level filtering on the performance of the DNF module. The uniform sam-

Table 4
Effects of the backbone model of the agent. 6 popular backbone models are tested, while their performance is shown in different rows.

Model	dis_{best}	dis_{ave}	*MD*-3(%)	*MD*-5(%)	*MD*-10(%)	1-*MD*-10(%)
R2D	2.67 ± 0.10	7.55 ± 0.14	77.23 ± 1.44	85.45 ± 1.35	93.36 ± 0.83	6.64 ± 0.83
R(2+1)D	3.62 ± 0.13	$\mathbf{7.36 \pm 0.13}$	70.04 ± 2.49	79.12 ± 2.13	88.13 ± 1.29	11.87 ± 1.29
C3D	2.23 ± 0.13	8.74 ± 0.27	82.15 ± 0.84	90.21 ± 0.76	95.97 ± 0.80	4.03 ± 0.80
P3D	2.27 ± 0.85	8.09 ± 0.42	80.66 ± 0.58	88.60 ± 0.78	94.47 ± 0.85	5.53 ± 0.85
ResNet+BiLSTM	2.08 ± 0.04	8.34 ± 0.19	81.77 ± 0.68	89.87 ± 0.66	96.54 ± 0.51	3.46 ± 0.51
C3D+BiLSTM	$\mathbf{1.97 \pm 0.10}$	8.69 ± 0.16	$\mathbf{83.22 \pm 0.46}$	$\mathbf{91.40 \pm 0.57}$	$\mathbf{97.44 \pm 0.85}$	$\mathbf{2.56 \pm 0.85}$

Table 5
Performance of the DNF module using different filtering methods. The second column (contains GT) indicates the percentage of output that contains the GT keyframe, while the third shows the opposite. The fourth column shows the averaged minimal distance between the output and the GT keyframe.

Model	contains GT (%)	without GT (%)	dis_{best}
uniform sampling	48.05	51.95	1.26
vanilla	83.78	16.22	5.37
window-level	89.57	10.43	3.63
dual-level	**95.75**	**4.25**	1.08

pling (down-/up-sampling all videos to the same length) only obtained 48.05% videos that contain the GT keyframe. Using the vanilla DNF (without any filtering), 83.78% output videos contain the GT frame. This is expected as the detection model could help to remove frames without lesions. However, the dis_{best} increased as the detection model is 2D and was trained using 2D frames. As a result, it might have problems handling frames with an ambiguous boundary or misleading surrounding tissues and produce unstable detection results (e.g. Fig. 5). Equipped with the window-level filtering, the DNF is able to smooth the prediction and produce more stable results (89.56% contains the GT). The proposed dual-level filtering used additional video-level filtering which takes information from the whole video into consideration and further boosts the performance (6.18% more than the window-level, and 11.97% more than the vanilla version). Note that even though 4.25% of the outputs do not contain the GT frame, they are still close to the GT ($dis_{best} = 1.08$).

5.5. Efficacy of the group-aware focal loss

Table 6 compares different loss function-based approaches in handling attribute classification facing severe class imbalance. The standard cross-entropy loss obtained relatively high Accuracy (80.16%) and Specificity (91.43%), while the Recall (37.36%) and Precision (63.00%) are substantially lower. This indicates a strong bias towards the negative class (i.e., the majority group). As a result, it scored poorly in F1-score and G-mean metrics, which consider the performance of both the majority and the minority groups. Focal loss obtained even higher Accuracy and Specificity, while the Recall is still low. We speculate that this was caused by the additional parameter used by focal loss, which decays the loss of easy examples (high prediction confidence) but does not guarantee a remedy for class imbalance. Therefore, the gradient of the majority group might still dominate the training, which could help to reduce false-positive predictions but still hamper the learning of the minority group. The ASL loss, on the other hand, scored higher on Recall (87.15%) while the Accuracy (80.16%) and Specificity (91.43%) decreased dramatically. Note that its F1-score (52.89%) and G-mean (63.12%) are higher than that of the Focal and cross-entropy loss, exhibiting superiority in handling imbalanced classes. This demonstrates that it is beneficial to treat the minority group differently, especially to increase the true positive rate. However, the ASL loss shifts the distribution of the prediction score by subtraction of 0.2 for the majority class. This reduces the influence of easy, negative examples, while also diminishing the study of hard, negative examples and may increase the false-positive rate (low Precision and Specificity). Meanwhile, the proposed GAFL loss also decouples the learning of the majority and the minority group while only cut-off the gradients of easy examples. It is able to achieve the highest F1-score (60.75%), G-mean (73.88%), and AUC (84.15%). It also shows a more balanced performance across Precision, Specificity, and Recall and is not biased towards certain groups. This suggests that the proposed GAFL can handle the extremely imbalanced data and is suitable for solving attribute classification of the breast lesion.

It should be noted that hyper-parameters (i.e., $\gamma_{-,+}$ and $th_{-,+}$) could influence the performance as well. Table 7 displays the hyper-parameter selection process for the GAFL loss. $th_{-,+}$ are first fixed to test the influence of $\gamma_{-,+}$ (row 1–3). γ_- are set with higher values to decrease the influence of easy, negative examples. It can be seen that the decoupling of the two groups (setting different values for γ_- and γ_+) has a large impact on the F1-score and G-mean. Note that the first row represents a loss that is essentially similar to the Focal loss ($\gamma_- = \gamma_+$). While Table 6 reports slightly higher results as we performed hyper-parameter tuning in the previous experiment for a fair comparison between CE, Focal, ASL, and the proposed GAFL loss (details reported in Section 4.2). The $\gamma_- = 3, \gamma_+ = 1$ is preferred for higher performance in F1-score and G-mean. Furthermore, to test the previous conjecture that γ_- should be set with a higher value than that of γ_+ to compensate for the influence of easy examples in the majority group, we carried out an additional experiment by setting $\gamma_- = 1, \gamma_+ = 3$. Results show that both F1-score and the G-mean drop significantly when reversing the value of $\gamma_{-,+}$. It also leads to a substantial decrease in Recall, while the Precision and Specificity are higher. This suggests the existence of a large number of false-negative predictions and a small number of false-positive predictions. The former, we argue, is especially undesirable in this circumstance as it indicates incompetence in recognizing the malignancy or malignancy indicators. We set $\gamma_- = 3, \gamma_+ = 1$ for all further experiments.

We then compare different sets of $th_{-,+}$ to examine their impact (row 5–10, Table 7). Similarly, th_- of higher values helps to reduce the effect of the easy, negative examples and scored higher F1-score and G-mean (row 5–7 compare to row 9, 10). However, an too aggressive setting of $th_{-,+}$ could also lead to a performance drop as it may interrupt the learning of normal cases in both groups. Overall, results prove that this probability cut-off can also alleviate the class imbalanced problem, while a proper hyper-parameter selection should be carried out for optimal performance. Note that all tested variants of GAFL scored higher F1-score and G-mean than the competing methods reported in Table.6. $th_- = 0.20, th_+ = 0.15$ are selected as its obtained the highest AUC, F1-score and G-mean. These parameters are used to train the ACN in the proposed KE-BUV framework.

Table 6
Comparison experiment for attribute classification. The cross-entropy (CE), Focal, ASL and the proposed GAFL loss are compared. Best results are highlighted in bold.

Loss	AUC (%)	Accuracy (%)	Recall (%)	Precision (%)	Specificity (%)	F1-score (%)	G-mean (%)
CE	80.40 ± 0.18	80.16 ± 0.42	37.36 ± 4.34	63.00 ± 2.31	91.43 ± 2.39	43.17 ± 2.90	53.57 ± 3.17
Focal	82.49 ± 0.06	**81.13 ± 0.11**	38.63 ± 7.16	**65.87 ± 2.68**	**92.83 ± 3.13**	47.88 ± 3.04	58.10 ± 2.95
ASL	80.22 ± 0.22	61.69 ± 2.98	**87.15 ± 2.78**	38.25 ± 1.72	49.43 ± 4.78	52.89 ± 1.22	63.12 ± 2.32
Ours	**84.15 ± 0.10**	78.34 ± 1.76	70.39 ± 6.21	53.99 ± 3.28	78.99 ± 4.65	**60.75 ± 0.44**	**73.88 ± 1.23**

Table 7
Hyper-parameter selection experiment for the GBAL loss. The value for $\gamma_{-,+}$ are first selected (row 1–3), then the $th_{-,+}$ are selected (row 4–7).

Parameter				AUC (%)	Accuracy (%)	Recall (%)	Precision (%)	Specificity (%)	F1-score (%)	G-mean (%)
γ_-	γ_+	th_-	th_+							
1	1	0	0	82.07 ± 0.21	80.55 ± 0.54	32.58 ± 5.92	64.25 ± 6.06	94.03 ± 2.67	39.21 ± 4.47	47.99 ± 4.35
2	1	0	0	83.19 ± 0.17	**80.70 ± 0.75**	53.76 ± 3.76	60.88 ± 2.32	87.39 ± 2.78	55.30 ± 0.87	66.53 ± 1.11
3	1	0	0	82.55 ± 0.22	77.32 ± 1.81	**67.07 ± 6.36**	52.49 ± 2.86	78.10 ± 4.57	**58.38 ± 1.00**	**71.26 ± 0.18**
1	3	0	0	**84.56 ± 0.16**	80.50 ± 0.30	22.38 ± 1.34	**79.22 ± 4.33**	**98.13 ± 0.26**	29.99 ± 1.25	37.25 ± 1.49
3	1	0.10	0.05	83.91 ± 0.12	78.95 ± 1.64	67.54 ± 5.97	55.23 ± 3.22	80.62 ± 4.17	60.35 ± 0.62	72.98 ± 1.51
3	1	0.15	0.10	83.42 ± 0.05	78.21 ± 2.95	67.08 ± 7.78	54.51 ± 4.48	79.89 ± 6.64	59.56 ± 2.67	72.34 ± 0.87
3	1	0.20	0.15	**84.15 ± 0.10**	78.34 ± 1.76	**70.39 ± 6.21**	53.99 ± 3.28	78.99 ± 4.65	**60.75 ± 0.44**	**73.88 ± 1.23**
3	1	0.25	0.20	83.36 ± 0.34	79.98 ± 1.09	61.74 ± 4.70	57.50 ± 2.66	84.28 ± 3.03	59.26 ± 0.84	71.43 ± 1.46
3	1	0.15	0.20	83.65 ± 0.34	80.00 ± 0.94	63.42 ± 4.14	57.59 ± 2.30	83.64 ± 0.26	60.07 ± 0.66	72.13 ± 1.18
3	1	0.10	0.15	84.08 ± 0.13	**80.30 ± 0.98**	63.10 ± 4.09	**58.07 ± 2.69**	**84.09 ± 2.82**	60.09 ± 0.55	72.04 ± 1.20

Table 8
Quality rating results of the automated generated keyframes. Doctors with difference experience rated the predicted keyframe sets into three categories based on their quality.

Group	Doctor	Excellent (no.)	Fair (no.)	Poor (no.)	Accept rate (%)
Junior	1	37	11	2	96
	2	40	8	2	96
	3	35	14	1	98
Senior	4	38	8	4	92
	5	37	13	0	100
	6	39	9	2	96
Ave	-	37.6	10.5	1.83	96.3

5.6. Reader study

Quality rating We first evaluate whether the automated generated keyframes are qualitative satisfactory to doctors. Table 8 reports the results. It can be seen that there is no significant difference between the two groups and most cases are marked with 'Excellent'. We further calculate the 'accept rate' based on the percentage of cases rated with 'Excellent' and 'Fair'. The average performance suggests that the proposed framework can produce automated keyframe sets that satisfy the need of clinicians (accept rate $\geq$ 96%). It is also worthwhile to examine when the proposed framework fails. Fig. 8 presents three cases that have been labeled as 'poor'. It can be seen that the automated generated keyframe sets are visually similar to the manually selected GT. These cases either have a small lesion (e.g., row 1 and 2 in Fig. 8) or an extremely obscure lesion margin (e.g., row 3). One of the reasons that the former examples were labeled as poor is that the selected keyframes are similar to each other. This is expected as the lesions only appear in a few frames within the video. As a result, the proposed framework picked these adjacent frames that inevitably share similar features. Another solution might be to explicitly constrain the keyframes to be different in the future. For lesions with obscure margins or occluded appearance, it might be difficult for both the clinicians and the automated framework to recognize the lesions and pick out suitable keyframes. The better choice may be to manipulate the probe to suitable angles and scan the lesion again for clearer visualization.

Cancer diagnosis Results are reported in Table 9. It can be seen that both groups obtained relatively lower accuracy using the static 2D images. This is expected as this single image might not contain enough information regarding the whole lesion. Using the keyframe sets, both junior and senior doctors achieved better diagnosis performance (compare row 1 to 4, row 2 to 5 in Table 9). This suggests that the automated generated keyframe sets contain comprehensive information about the lesion, and could assist the diagnosis. Meanwhile, the junior doctors benefit more from the keyframe set (e.g., a 13.33% increase in accuracy, a 7.14% increase in F1-score) compared to that of the senior group. This may be explained by that junior doctors need more context to evaluate a lesion while the senior doctors have higher experience levels and they can perform relatively accurate diagnoses with limited information.

Given the full US breast videos, both groups scored inferiorly to that obtained from the keyframe sets (compare row 4 to 7, row 5 to 8 in Table 9). This may result from that there may exist large redundant information and misleading tissues in breast US videos that could cause distraction. These results further help to prove the efficacy of the proposed framework. It is also interesting to see that the senior doctors score higher accuracy using 2D images than those using the whole breast videos. This may stem from the fact

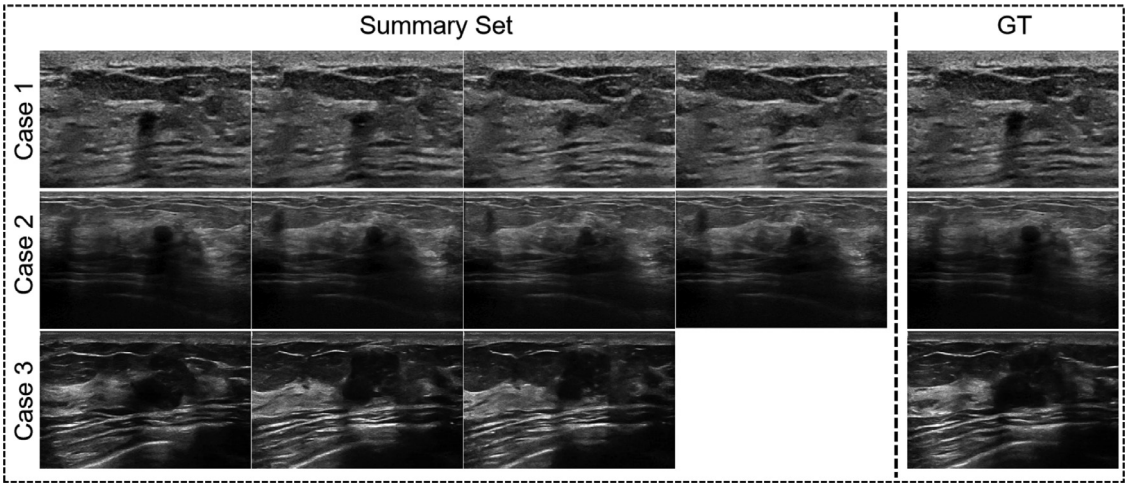

Fig. 8. Challenging cases. These cases are rated as 'poor' in the quality rating study. The automated generated keyframe sets are shown on the left, while the corresponding GT keyframes are shown on the right.

Table 9
Diagnosis performance of doctors with different experience levels given different inputs (i.e. 2D US image, the keyframe set, and the full US video).

Input	Reader	AUC (%)	Accuracy (%)	Recall (%)	Precision (%)	Specificity(%)	F1-score (%)
2D Image	Junior	67.14 ± 5.30	62.00 ± 10.58	44.73 ± 9.56	80.00 ± 13.33	87.15 ± 5.74	56.27 ± 4.83
	Senior	83.97 ± 4.82	82.00 ± 6.00	65.45 ± 8.90	88.89 ± 7.70	94.46 ± 3.35	75.03 ± 6.49
	Ave	75.56 ± 10.27	72.00 ± 13.39	55.09 ± 14.04	84.44 ± 10.88	90.81 ± 5.81	65.65 ± 11.48
Auto Keyframe Set	Junior	73.97 ± 5.05	73.33 ± 8.08	55.65 ± 9.79	75.56 ± 7.70	87.42 ± 2.39	63.41 ± 5.80
	Senior	84.60 ± 3.82	84.67 ± 1.15	71.18 ± 5.24	84.44 ± 13.88	93.25 ± 5.99	76.57 ± 2.91
	Ave	79.28 ± 7.07	79.00 ± 8.07	63.41 ± 11.03	80.00 ± 11.16	90.33 ± 5.18	69.99 ± 8.29
US Video	Junior	73.49 ± 2.44	78.00 ± 2.00	63.65 ± 3.38	62.22 ± 3.85	83.97 ± 1.53	62.92 ± 3.41
	Senior	81.75 ± 4.29	83.33 ± 2.31	70.11 ± 3.00	77.78 ± 10.18	90.16 ± 3.84	73.50 ± 4.95
	Ave	77.62 ± 5.50	80.67 ± 3.50	66.88 ± 4.55	70.00 ± 10.96	87.06 ± 4.28	68.21 ± 6.93

that these 2D images are not random frames extracted from the videos, but instead are the manually selected keyframes (one for each video). Senior doctors might be accustomed to making a diagnosis using the single selected keyframe and thus scored higher diagnosis accuracy. Note that for the purpose of this study, no other information regarding the patient (e.g., age, medical record) was given to the doctors, while in real-world scenarios they could possibly utilize this information to further assist the diagnosis.

6. Conclusion

In this paper, we proposed a novel keyframe extraction framework bespoke for US videos of breast lesions. It leverages the reinforcement learning technique to coordinate the learning between the primary keyframe extraction task and the auxiliary lesion detection and attribute classification tasks. It is also equipped with a dual-level filtering module to filter irrelevant information and avoid distraction. We also designed a group-aware Focal loss to combat the severe imbalanced data in attribute classification. The experimental results showed that our method can outperform other state-of-the-art algorithms in video summary or keyframe extraction. Ablation studies and additional experiments were also carried out to verify the efficacy of each component or examine the influence of different hyper-parameters. Finally, reader studies were performed to test whether the automated generated keyframe sets are acceptable by doctors with different experience levels and whether they could benefit subsequent diagnosis of breast lesions. In the future, the proposed framework may also be applied to analyze US videos of other organs or automated whole-breast US–given suitable adaption to process 3D images (e.g., 3D lesion detection model and 3D frame sampling method).

Declaration of Competing Interest

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62101342, and No. 62171290); Shenzhen-Hong Kong Joint Research Program (No. SGDX20201103095613036), Shenzhen Science and technology research and Development Fund for Sustainable development project (No. KCXFZ20201221173613036).; Key project of Regional Joint Fund of Guangdong Provincial Natural Science Foundation (No.2020B1515120098).

We would also like to thank: Ling Guan, Department of Ultrasound, Gansu Provincial Cancer Hospital, Lanzhou, China; Qunxia Zhang, Department of Ultrasound, the Second Affiliated Hospital of Chongqing Medical University, Chongqing, China; Yan Cai, Department of Ultrasound Medicine, Central Peoples Hospital of Zhanjiang, Zhanjiang, China; Miaoli Shi, Department of Ultrasound, Boai Hospital of Zhongshan City, Zhongshan, China; Xiaohong Xu, Department of Ultrasound, Affiliated Hospital of Guangdong Medical University, Zhanjiang, China; Li Chen, Department of Ultrasound, The First Affiliated Hospital of Nanchang University, Nanchang, China; Heqing Zhang and Yulan Peng, Department of Ultrasound, West China Hospital, Sichuan University, Chengdu, China; for their support to this study. We would like to acknowledge Pair[3] annotation software for its technical support.

[3] https://aipair.com.cn/en/

References

American College of Radiology and others, 2003. Breast Imaging Reporting and Data System-Ultrasound (bi-rads. American College of Radiology, Reston.

Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., Patras, I., 2021. Video summarization using deep neural networks: a survey. arXiv preprint arXiv:2101.06072.

Asha Paul, M.K., Kavitha, J., Jansi Rani, P.A., 2018. Key-frame extraction techniques: a review. Recent Pat. Comput. Sci. 11 (1), 3–16.

Baumgartner, C. F., Kamnitsas, K., Matthew, J., Smith, S., Kainz, B., Rueckert, D., 2016. Real-time standard scan plane detection and localisation in fetal ultrasound using fully convolutional neural networks. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 203–211.

Boyd, N.F., Martin, L.J., Bronskill, M., Yaffe, M.J., Duric, N., Minkin, S., 2010. Breast tissue composition and susceptibility to breast cancer. J. Natl. Cancer Inst. 102 (16), 1224–1237.

Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 106, 249–259.

Cai, Y., Droste, R., Sharma, H., Chatelain, P., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2020. Spatio-temporal visual attention modelling of standard biometry plane-finding navigation. Med. Image Anal. 65, 101762.

Chao, T.-C., Lo, Y.-F., Chen, S.-C., Chen, M.-F., 1999. Prospective sonographic study of 3093 breast tumors. J. Ultrasound Med. 18 (5), 363–370.

Chen, H., Wu, L., Dou, Q., Qin, J., Li, S., Cheng, J.-Z., Ni, D., Heng, P.-A., 2017. Ultrasound standard plane detection using a composite neural network framework. IEEE Trans. Cybern. 47 (6), 1576–1586.

Ciompi, F., Pujol, O., Balocco, S., Carrillo, X., Mauri-Ferré, J., Radeva, P., 2011. Automatic key frames detection in intravascular ultrasound sequences. Proceedings of the 14th MICCAI, 78–94.

Dou, H., Yang, X., Qian, J., Xue, W., Qin, H., Wang, X., Yu, L., Wang, S., Xiong, Y., Heng, P.-A., et al., 2019. Agent with warm start and active termination for plane localization in 3D ultrasound. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 290–298.

Fajtl, J., Sokeh, H. S., Argyriou, V., Monekosso, D., Remagnino, P., 2018. Summarizing videos with attention. In: Proceedings of the Asian Conference on Computer Vision. Springer, pp. 39–54.

Han, S., Kang, H.-K., Jeong, J.-Y., Park, M.-H., Kim, W., Bang, W.-C., Seong, Y.-K., 2017. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys. Med. Biol. 62 (19), 7714.

Heinig, J., Witteler, R., Schmitz, R., Kiesel, L., Steinhard, J., 2008. Accuracy of classification of breast ultrasound findings based on criteria used for bi-rads. Ultrasound Obstet. Gynecol. 32 (4), 573–578. The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology

Hernández-López, J., Gómez-Flores, W., 2020. Predicting the bi-rads lexicon for mammographie masses using hybrid neural models. In: Proceedings of the 17th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE). IEEE, pp. 1–6.

Huang, J.-H., Worring, M., 2020. Query-controllable video summarization. In: Proceedings of the International Conference on Multimedia Retrieval, pp. 242–250.

Huang, R., Lin, Z., Dou, H., Wang, J., Miao, J., Zhou, G., Jia, X., Xu, W., Mei, Z., Dong, Y., et al., 2021. Aw3m: an auto-weighting and recovery framework for breast cancer diagnosis using multi-modal ultrasound. Med. Image Anal. 102137.

Jiao, J., Cai, Y., Alsharid, M., Drukker, L., Papageorghiou, A. T., Noble, J. A., 2020. Self-supervised contrastive video-speech representation learning for ultrasound. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 534–543.

JM, S.M., JM, L.C., et al., 2009. Diagnostic accuracy and interobserver variability in the bi-rads ultrasound system. Radiologia 51 (5), 477–486. (Panama)

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6 (1), 1–54.

Kim, K.W., Cho, K.R., Seo, B.K., Whang, K.W., Woo, O.H., Oh, Y.W., Kim, Y.H., Bae, J.W., Park, Y.S., Hwang, C.M., et al., 2010. Sonographic findings of mammary duct ectasia: can malignancy be differentiated from benign disease? J Breast Cancer 13 (1), 19–26.

Li, B., Liu, Y., Wang, X., 2019. Gradient harmonized single-stage detector. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33, pp. 8577–8584.

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

Liu, T., Meng, Q., Vlontzos, A., Tan, J., Rueckert, D., Kainz, B., 2020a. Ultrasound video summarization using deep reinforcement learning. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 483–492.

Liu, Y., An, X., Cong, L., Dong, G., Zhu, L., 2020. Embedding weighted feature aggregation network with domain knowledge integration for breast ultrasound image segmentation. In: Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis. Springer, pp. 66–74.

Mahasseni, B., Lam, M., Todorovic, S., 2017. Unsupervised video summarization with adversarial lstm networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 202–211.

Mohammed, M.A., Al-Khateeb, B., Rashid, A.N., Ibrahim, D.A., Abd Ghani, M.K., Mostafa, S.A., 2018. Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images. Comput. Electr. Eng. 70, 871–882.

Organization, W. H., et al., 2020. Latest global cancer data: cancer burden rises to 19.3 million new cases and 10.0 million cancer deaths in 2020.

Otani, M., Nakashima, Y., Rahtu, E., Heikkila, J., 2019. Rethinking the evaluation of video summaries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7596–7604.

Pan, G., Zheng, Y., Zhang, R., Han, Z., Sun, D., Qu, X., 2019. A bottom-up summarization algorithm for videos in the wild. EURASIP J. Adv. Signal Process. 2019 (1), 1–11.

Parihar, A. S., Mittal, R., Jain, P., et al., 2021. Survey and comparison of video summarization techniques. In: Proceedings of the 5th International Conference on Computer, Communication and Signal Processing (ICCCSP). IEEE, pp. 268–272.

Paulinelli, R.R., Freitas-Júnior, R., Moreira, M.A.R., Alves de Moraes, V., Bernardes-Júnior, J.R.M., Vidal, C.d.S.R., Ruiz, A.N., Lucato, M.T., 2005. Risk of malignancy in solid breast nodules according to their sonographic features. J. Ultrasound Med. 24 (5), 635–641.

Pu, B., Li, K., Li, S., Zhu, N., 2021. Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT. IEEE Transactions on Industrial Informatics, 17(11),, 7771–7780.

Rahman, M. R., Shah, S., Subhlok, J., 2020. Visual summarization of lecture video segments for enhanced navigation. In: Proceedings of the IEEE International Symposium on Multimedia (ISM). IEEE, pp. 154–157.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. 1804.02767.

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L., 2021. Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 82–91.

Rochan, M., Ye, L., Wang, Y., 2018. Video summarization using fully convolutional sequence networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 347–363.

Senie, R.T., Rosen, P.P., Lesser, M.L., Kinne, D.W., 1981. Breast self-examination and medical examination related to breast cancer stage. Am. J. Public Health 71 (6), 583–590.

Shen, S., Zhou, Y., Xu, Y., Zhang, B., Duan, X., Huang, R., Li, B., Shi, Y., Shao, Z., Liao, H., et al., 2015. A multi-centre randomised trial comparing ultrasound vs mammography for screening breast cancer in high-risk chinese women. Br. J. Cancer 112 (6), 998–1004.

Stoean, R., Iliescu, D., Stoean, C., Ilie, V., Patru, C., Hotoleanu, M., Nagy, R., Ruican, D., Trocan, R., Marcu, A., et al., 2021. Deep learning for the detection of frames of interest in fetal heart assessment from first trimester ultrasound. In: Proceedings of the International Work-Conference on Artificial Neural Networks. Springer, pp. 3–14.

Ting, F.F., Tan, Y.J., Sim, K.S., 2019. Convolutional neural network improvement for breast cancer classification. Expert Syst. Appl. 120, 103–115.

Van Hulse, J., Khoshgoftaar, T. M., Napolitano, A., 2007. Experimental perspectives on learning from imbalanced data. In: Proceedings of the 24th International Conference on Machine Learning, pp. 935–942.

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., Kennedy, P. J., 2016. Training deep neural networks on imbalanced data sets. IEEE, pp. 4368–4374.

Wang, J., Miao, J., Yang, X., Li, R., Zhou, G., Huang, Y., Lin, Z., Xue, W., Jia, X., Zhou, J., Huang, R., Ni, D., 2020. Auto-weighting for breast cancer classification in multimodal ultrasound. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Springer International Publishing, Cham, pp. 190–199.

Williams, R.J., 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. 8 (3–4), 229–256.

Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D., 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 162–178.

Yan, X., Gilani, S. Z., Qin, H., Feng, M., Zhang, L., Mian, A., 2018. Deep keyframe detection in human action videos. arXiv preprint arXiv:1804.10021.

Yap, B. W., Abd Rani, K., Abd Rahman, H. A., Fong, S., Khairudin, Z., Abdullah, N. N., 2014. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In: Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). Springer, pp. 13–22.

Zhang, E., Seiler, S., Chen, M., Lu, W., Gu, X., 2020. Birads features-oriented semi–supervised deep learning for breast ultrasound computer-aided diagnosis. Phys. Med. Biol. 65 (12), 125005.

Zhang, K., Chao, W.-L., Sha, F., Grauman, K., 2016. Video summarization with long short-term memory. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 766–782.

Zhou, K., Qiao, Y., Xiang, T., 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: Proceedings of the AAAI Conference on Artificial Intelligence 32 (1).