Boosting Breast Ultrasound Video Classification by the Guidance of Keyframe Feature Centers

Anlan Sun[1], Zhao Zhang[2], Meng Lei[2], Yuting Dai[1], Dong Wang[3], and Liwei Wang[3,4(✉)]

[1] Yizhun Medical AI Co., Ltd., Beijing, China
{anlan.sun,yuting.dai}@yizhun-ai.com
[2] Center for Data Science, Peking University, Beijing, China
{zhangzh,leimeng}@stu.pku.edu.cn
[3] National Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University, Beijing, China
{wangdongcis,wanglw}@pku.edu.cn
[4] Center for Machine Learning Research, Peking University, Beijing, China

Abstract. Breast ultrasound videos contain richer information than ultrasound images, therefore it is more meaningful to develop video models for this diagnosis task. However, the collection of ultrasound video datasets is much harder. In this paper, we explore the feasibility of enhancing the performance of ultrasound video classification using the static image dataset. To this end, we propose KGA-Net and coherence loss. The KGA-Net adopts both video clips and static images to train the network. The coherence loss uses the feature centers generated by the static images to guide the frame attention in the video model. Our KGA-Net boosts the performance on the public BUSV dataset by a large margin. The visualization results of frame attention prove the explainability of our method. We release the code and model weights in https://github.com/PlayerSAL/KGA-Net.

Keywords: Breast ultrasound classification · Ultrasound video · Coherence loss

1 Introduction

Breast cancer is a life-threatening disease that has surpassed lung cancer as leading cancer in some countries and regions [20]. Breast ultrasound is the primary screening method for diagnosing breast cancer, and accurately distinguishing between malignant and benign breast lesions is crucial. This task is also an essential component of computer-aided diagnosis. Since each frame in an ultrasound video can only capture a specific view of a lesion, it is essential to aggregate information from the entire video to perform accurate automatic lesion diagnosis. Therefore, in this study, we focus on the classification of breast ultrasound videos for detecting malignant and benign breast lesions.

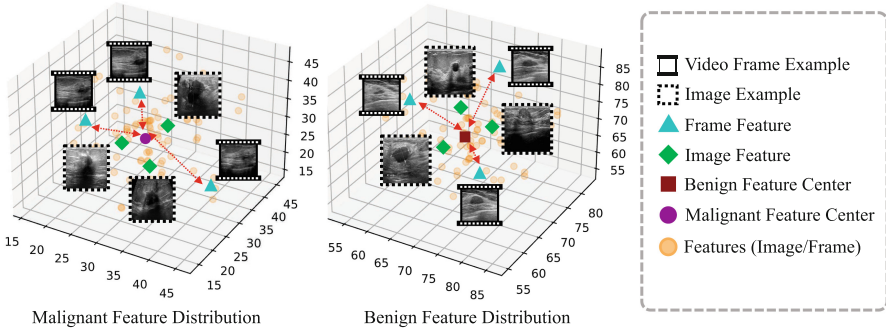A. Sun and Z. Zhang—Equal contribution.

Fig. 1. Feature distribution of video frames from BUSV [15] and static images from BUSI [1]. We use a 2D ResNet trained on ultrasound images to get the features.

While ultrasound videos offer more information, prior studies have primarily focused on static image classification [2,11,27]. Obtaining ultrasound video data with pathology gold standard results poses a major challenge. Sonographers typically record keyframe images during general ultrasound examinations, not entire videos. Prospective collection requires additional efforts to track corresponding pathological results. Consequently, while there are many breast ultrasound image datasets [1,28], breast ultrasound video datasets remain scarce, with only one relatively small dataset [15] containing 188 videos available currently.

Given the difficulties in collecting ultrasound video data, we investigate the feasibility of enhancing the performance of ultrasound video classification using a static image dataset. To achieve this, we first analyze the relationship between ultrasound videos and images. The images in the ultrasound dataset are keyframes of a lesion that exhibit the clearest appearance and most typical symptoms, making them more discriminative for diagnosis. Although ultrasound videos provide more information, the abundance of frames may introduce redundancy or vagueness that could disrupt classification. From the aspect of feature distribution, as shown in Fig. 1, the feature points of static images are more concentrated, while the feature of video frames sometimes are away from the class centers. Frames far from the centers are harder to classify. Therefore, it is a promising approach to guide the video model to pay more attention to important frames close to the class center with the assistance of static keyframe images. Meanwhile, our approach aligns with the diagnosis of ultrasound physicians, automatically evaluates the importance of frames, and diagnoses based on the information of key frames. Additionally, our method provides interpretability through key frames.

In this paper, we propose a novel Keyframe Guided Attention Network (KGA-Net) to boost ultrasound video classification. Our approach leverages both image (keyframes) and video datasets to train the network. To classify videos, we use frame attention to predict feature weights for all frames and aggregate them to make the final classification. The feature weights determine the contribution

of each frame for the final diagnosis. During training, we construct category feature centers for malignant and benign examples respectively using center loss [26] on static image inputs and use the centers to guide the training of video frame attention. Specifically, we propose coherence loss, which promotes the frames close to the centers to have high attention weights and decreases the weights for frames far from the centers. Due to the feature centers being generated by the larger scale image dataset, it provides more accurate and discriminative feature centers which can guide the video frame attention to focus on important frames, and finally leads to better video classification.

Our experimental results on the public BUSV dataset [15] show that our KGA-Net significantly outperforms other video classification models by using an external ultrasound image dataset. Additionally, we visualized attention values guided by the coherence loss. The frames with clear diagnostic characteristics are given higher attention values. This phenomenon makes our method more explainable and provides a new perspective for selecting keyframes from video.

In conclusion, our contributions are as follows:

1. We analyze the relationship between ultrasound video data and image data, and propose the coherence loss to use image feature centers to guide the training of frame attention.
2. We propose KGA-Net, which adopts a static image dataset to boost the performance of ultrasound video classification. KGA-Net significantly outperforms other video baselines on the BUSV dataset.
3. The qualitative analysis of the frame attention verifies the explainability of our method and provides a new perspective for selecting keyframes.

2   Related Works

Breast Ultrasound Classification. Breast ultrasound (BUS) plays an important supporting role in the diagnosis of breast-related diseases. Recent research demonstrated the potential of deep learning for breast lesion classification tasks [6,18,19,23,27]. [6,18] design ensemble methods to integrate the features of multiple models to obtain higher accuracy. [19,23,27] utilize multi-task learning to improve the model performance. However, all of them are based on image datasets, such as BUSI [1], while few works focus on the video modality. [14] design a pre-training model based on contrastive learning for ultrasound video classification. [13,25] develop a keyframe extraction model for ultrasound videos and utilized the extracted keyframes to perform various classification tasks. However, these methods rely on keyframe supervision, which limits their applicability. Fortunately, the recent publicly available dataset BUSV [15] has made the research on the task of BUS video-based classification possible. In this paper, we build our model based on this dataset.

Video Recognition Based on Neural Networks. Traditional methods are based on Two-stream networks [9,10,24]. Since I3D [3] was proposed, 3D CNNs have dominated video understanding for a long time. [21,22] decompose 3D convolution in different ways to reduce computation complexity without losing
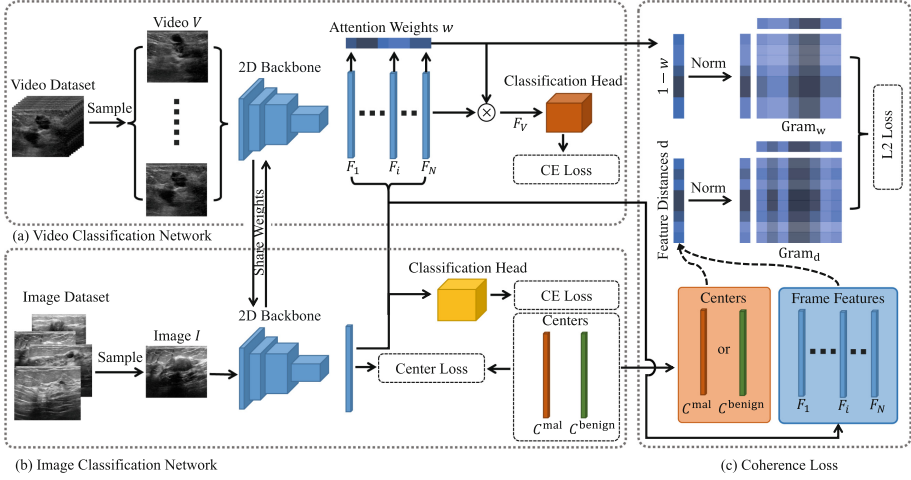
Fig. 2. Overview of our proposed keyframe-guided attention network.

performance. [8] designed two branches to focus on temporal information and spatial features, respectively. However, 3D CNNs have a limited receptive field, and thus struggle to capture long-range dependency. Vision Transformers [5,16] have become popular for their ability to aggregate spatial-temporal information. To address computational complexity, MViT [7] employed a hierarchical structure and Video Swin [17] introduced 3D shifted window attention. Our proposed KGA-Net is a simple framework that leverages the frame attention module to aggregate multi-frame features efficiently.

3   Methodology

As shown in Fig. 2, our KGA-Net takes the video inputs and static image inputs simultaneously to train the network. The coherence loss is proposed to guide the frame attention by using the feature centers generated by the images. We will then elaborate on each component in the following sections.

3.1   Video and Image Classification Network

The video classification network is illustrated in Fig. 2 (a). The model is composed of a 2D CNN backbone, a frame attention module, and a classification head. For an input video clip V composed of N frames, it is first processed by the backbone network and the feature vectors of the frames $\{F_i\}_{i=1}^{N}$ are obtained. Then, the frame attention module predicts the attention weight for each frame using a FC and sigmoid layer, and then the features are aggregated by the weights to form an integrated feature vector. Formally,

$$w_i = \text{Sigmoid}(\text{FC}(F_i)) \tag{1}$$

where w_i denotes the weight for the i_{th} frame and FC is the fully-connected layer. Then, the features are aggregated by $F_V = \sum_{i=1}^{N} w_i \cdot F_i$. Finally, the classification head is applied to the final result of lesion classification. To train the model, the cross-entropy loss (CE Loss) is applied to the classification prediction of the video.

The image classification network is used to assist in training the video model. We use the same 2D CNN as the backbone network in the video classification network. The model weights are shared for the two backbones for better generalization. To promote the formation of feature centers, we apply the center loss [26] to the image model besides the cross-entropy loss. In addition, the frame-level cross-entropy loss is also applied to the video frames to facilitate training.

3.2   Training with Coherence Loss

In this section, we introduce the coherence loss to guide the frame attention with the assistance of the category feature centers. We use the same method as center loss [26] to obtain the feature centers for the malignancy and benign lesions, which are denoted as $\mathcal{C}^{\text{mal}}$ and $\mathcal{C}^{\text{benign}}$, respectively.

The distances of frame features and the feature centers can measure the quality of the frames. The frame features close to the centers are more discriminative for the classification task. Therefore, we use these distances to guide the generation of frame attention. Specifically, we push the frames close to the centers to have higher attention weights and decrease the weights far from the centers. To do this, for each video frame with feature F_i, we first calculate the feature distance from its corresponding class center. Formally,

$$d_i = \|F_i - \mathcal{C}^Y\|_2, \tag{2}$$

where $Y \in \{\text{mal}, \text{benign}\}$ is the label of the video V and d_i is the computed distance of frame i.

Afterward, we apply coherence loss to the attention weights $\mathbf{w} = [w_1, w_2, ..., w_N]^\intercal$ to make them have a similar distribution with the feature distances $\mathbf{d} = [d_1, d_2, ..., d_N]^\intercal$. To supervise the distribution, the coherence loss is defined as the L2 loss of the gram matrix of these two vectors

$$\text{L}_{\text{Coh}} = \|\text{Gram}_{\mathbf{w}} - \text{Gram}_{\mathbf{d}}\|_2, \tag{3}$$

where $\text{Gram}_{\mathbf{w}} = \frac{(\mathbf{1}-\mathbf{w}) \cdot (\mathbf{1}-\mathbf{w})^\intercal}{\|\mathbf{1}-\mathbf{w}\|_2^2}$ is the gram matrix of normalized attention weights, and $\text{Gram}_{\mathbf{d}} = \frac{\mathbf{d} \cdot \mathbf{d}^\intercal}{\|\mathbf{d}\|_2^2}$ is the gram matrix of normalized feature distances. Note that lower distances correspond to stronger attention, hence we use the opposite of $\mathbf{w}$ to get $\text{Gram}_{\mathbf{w}}$.

3.3   Total Training Loss

To summarize, the total training loss of our KGA-Net

$$L_{\text{total}} = L_{\text{CE}}^V + L_{\text{CE}}^I + L_{\text{Center}} + \lambda \cdot L_{\text{Coh}}. \tag{4}$$

L_{CE}^V and L_{CE}^I denote the cross-entropy for video classification and image and frame classification. L_{Center} means the center loss. λ is the weight for coherence loss. Empirically, we set $\lambda = 1$ in our experiments.

During inference, to perform classification on video data, the video classification network can be utilized individually for prediction.

4   Experiments

4.1   Implementation Details

Datasets. We use the public BUSV dataset [15] for video classification and the BUSI dataset [1] as the image dataset. BUSV consists of 113 malignant videos and 75 benign videos. BUSI contains 445 images of benign lesions and 210 images of malignant lesions. For the BUSV dataset, we use the official data split in [15]. All images of the BUSI dataset are adopted to train our KGA-Net.
Model Details. ResNet-50 [12] pretrained on ImageNet [4] is used as backbone. We use SGD optimizer with an initial learning rate of 0.005, which is reduced by $10\times$ at the 4,000th and 6,000th iteration. The total learning iteration number is 8,000. The learning rate warmup is used in the first 1,000 iterations. For each batch, the video clips and static images are both sampled and sent to the network. We use a total batchsize of 16 and the sample probability of video clips and images is 1:1. We implement the model based on Pytorch and train it with NVIDIA Titan RTX GPU cards.

During inference, we use the video classification network individually. In order to satisfy the fixed video length requirement of MViT [7], we sample up to 128 frames of each video to form a video clip and predict its classification result using all the models in experiments.

4.2   Comparison with Video Models

In this section, we compare our KGA-Net with other competitive video classification models. However, comparing with ultrasound-video-based work is challenging due to limited code accessibility and lack of keyframe detection model in existing methods [13,14,25]. Therefore, we compare our method with strong video baselines on natural images. We include CNN-based models, I3D [3], Slow-Fast [8], R(2+1)D [22], and CSN [21], along with the popular transformer-based model MViT [7]. For fairness comparison, we train these models using both video and image data, treating images as static videos. Evaluation metrics are reported on the BUSV test set for performance assessment.

As shown in Table 1, by leveraging the guidance of the image dataset, our KGA-Net significantly surpasses all other models on all of the metrics. The video classification model of our KGA-Net is composed of a standard 2D ResNet-50 and a light feature attention module, while the baseline models are with net structures carefully designed for video analysis. Therefore, the success of our KGA-Net lies in the correct usage of the image guidance. The feature centers

Table 1. Comparison with other video models. Classification thresholds are determined by Youden index.

Model	AUC(%)	ACC(%)	Sensitivity(%)	Specificity(%)
I3D [3]	88.31	81.58	84.00	76.92
SlowFast [8]	82.54	79.49	76.92	84.62
R(2+1)D [22]	86.46	81.58	84.00	76.92
CSN [21]	83.38	81.58	84.00	76.92
MViT [7]	90.53	82.05	80.77	84.62
KGA-Net (Our)	**94.67**	**89.74**	**88.46**	**92.31**

formed by the image dataset with larger data size and clear appearance effectively improve the accuracy of frame attention hence boosting the video classification performance.

4.3 Ablation Study

In this section, we ablate the contribution of each key design in our KGA-Net. We observe their importance by removing these key components from the whole network. The results are shown in Table 2. The results of KGA-Net are shown in the last row in Table 2, while the components are ablated in the first three rows. We use the same training schedule for all of the experiments.

Image guidance is the main purpose of our method. To portray the effect of using the image dataset, we train the KGA-Net using BUSV dataset alone in the first row of Table 2. Without the image dataset, we generate the feature centers from the video frames. As a result, the performance significantly drops due to the decrease in dataset scale. It also shows that the feature centers generated by the image dataset are more discriminative than that of the video dataset. It is not only because the lesion number of BUSI is larger than BUSV, but also because the images in BUSI are all the keyframes that contain typical characteristics of lesions.

Frame attention and coherence loss are two essential modules of our KGA-Net. We train a KGA-Net without the coherence loss in the third row of Table 2. In the second row, we further replace the feature attention module with feature averaging of video frames. It can be seen that both of these two modules contribute to the overall performance according to AUC and ACC. It is worth noting that these two models without coherence loss obtain very low sensitivity and high specificity, which means the model predictions are imbalanced and intend to make benign predictions. It is because that clear malignant appearances usually only exist in limited frames in a malignant video. Without our coherence loss or frame attention, it is difficult for the model to focus on typical frames that possess malignant features. This phenomenon certifies the effectiveness of our KGA-Net to prevent false negatives in diagnosis.

Table 2. Ablation studies. Model components are removed in the first three lines to analyze their contributions in KGA-Net. Classification thresholds are determined by Youden index.

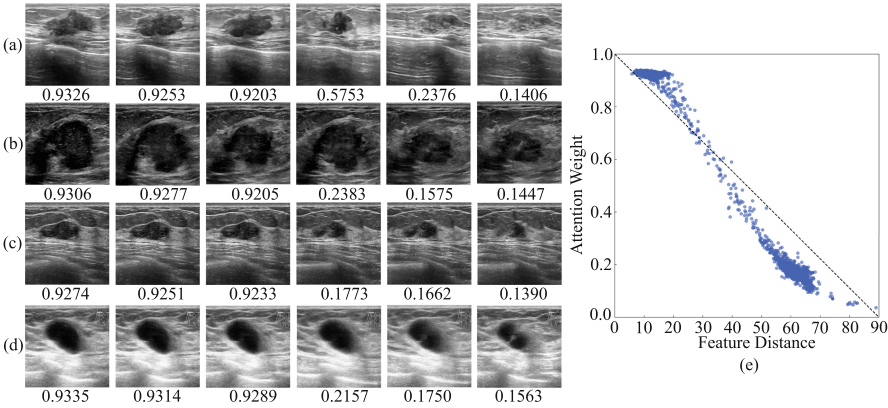| Model	AUC(%)	ACC(%)	Sensitivity(%)	Specificity(%)
w/o image guidance	85.21	76.92	73.08	84.62
w/o coherence loss & attention	88.17	74.36	61.54	**100.0**
w/o coherence loss	92.90	87.18	80.77	**100.0**
KGA-Net	**94.67**	**89.74**	**88.46**	92.31

4.4   Visual Analysis

Fig. 3. Visual Analysis. (a–d) Visualization of video frames and corresponding frame attention weights. (e) Relationship between attention weight and feature distance.

In Fig. 3, we illustrate video frames with their corresponding frame attention weights predicted by KGA-Net. Overall speaking, the frames with high attention weights do have clear image appearances for diagnosis. For example, the first three frames in Fig. 3(b) clearly demonstrate the edge micro-lobulation and irregular shapes, which lead to malignant judgment. Furthermore, we plot the relationships between the predicted attention values and the feature distances to the centers. As shown in Fig. 3(e), these two variables are linearly related, which indicates that KGA-Net the attention weights are effectively guided by the feature distances.

The qualitative analysis proves the interpretability of our method, which will benefit clinical usage. Moreover, the attention weights reveal the importance of each frame for lesion diagnosis. Therefore, it can provide a new perspective for the keyframe extraction task of ultrasound videos.

5   Conclusion

We propose KGA-Net, a novel video classification model for breast ultrasound diagnosis. Our KGA-Net takes as input both the video data and image data to train the network. We propose the coherence loss to guide the training of the video model by the guidance of feature centers of the images. Our method significantly exceeds the performance of other competitive video baselines. The visualization of the attention weights validates the effectiveness and interpretability of our KGA-Net.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020)
2. Byra, M.: Breast mass classification with transfer learning based on scaling of deep representations. Biomed. Signal Process. Control **69**, 102828 (2021)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
5. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Eroğlu, Y., Yildirim, M., Çinar, A.: Convolutional neural networks based classification of breast ultrasonography images by hybrid method with respect to benign, malignant, and normal using mRMR. Comput. Biol. Med. **133**, 104407 (2021)
7. Fan, H., et al.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6824–6835 (2021)
8. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019)
9. Feichtenhofer, C., Pinz, A., Wildes, R.: Spatiotemporal residual networks for video action recognition. In: Advances in Neural Information Processing Systems (NIPS), pp. 3468–3476 (2016)
10. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
11. Gheflati, B., Rivaz, H.: Vision transformers for classification of breast ultrasound images. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 480–483. IEEE (2022)

12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Huang, R., et al.: Extracting keyframes of breast ultrasound video using deep reinforcement learning. Med. Image Anal. **80**, 102490 (2022)
14. Lin, Z., Huang, R., Ni, D., Wu, J., Luo, B.: Masked video modeling with correlation-aware contrastive learning for breast cancer diagnosis in ultrasound. In: Xu, X., Li, X., Mahapatra, D., Cheng, L., Petitjean, C., Fu, H. (eds.) Resource-Efficient Medical Image Analysis: First MICCAI Workshop, REMIA 2022, Singapore, 22 September 2022, Proceedings, pp. 105–114. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16876-5_11
15. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., Wang, L.: A new dataset and a baseline model for breast lesion detection in ultrasound videos. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part III, pp. 614–623. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_59
16. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
17. Liu, Z., et al.: Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3202–3211 (2022)
18. Moon, W.K., Lee, Y.W., Ke, H.H., Lee, S.H., Huang, C.S., Chang, R.F.: Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Comput. Methods Programs Biomed. **190**, 105361 (2020)
19. Podda, A.S., Balia, R., Barra, S., Carta, S., Fenu, G., Piano, L.: Fully-automated deep learning pipeline for segmentation and classification of breast ultrasound images. J. Comput. Sci. **63**, 101816 (2022)
20. Siegel, R.L., et al.: Colorectal cancer statistics, 2017. CA: Cancer J. Clin. **67**(3), 177–193 (2017)
21. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5552–5561 (2019)
22. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
23. Wang, J., et al.: Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation. Med. Image Anal. **83**, 102687 (2023)
24. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 20–36. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_2
25. Wang, Y., et al.: Key-frame guided network for thyroid nodule recognition using ultrasound videos. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part IV, pp. 238–247. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_23

26. Wen, Y., Zhang, K., Li, Z., Qiao, Yu.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi. org/10.1007/978-3-319-46478-7_31
27. Zhang, G., Zhao, K., Hong, Y., Qiu, X., Zhang, K., Wei, B.: SHA-MTL: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification. Int. J. Comput. Assist. Radiol. Surg. **16**, 1719–1725 (2021)
28. Zhang, Y., et al.: BUSIS: a benchmark for breast ultrasound image segmentation. In: Healthcare, vol. 10, p. 729. MDPI (2022)