Real-world Visual Navigation for Cardiac Ultrasound View Planning

Mingkun Bao[1], Yan Wang[1,4], Xinlong Wei[1], Bosen Jia[5], Xiaolin Fan[4], Dong Lu[1], Yifan Gu[4], Jian Cheng[1], Yingying Zhang[2](✉), Chuanyu Wang[3](✉), and Haogang Zhu[1,2](✉)

[1] State Key Laboratory of Complex & Critical Software Environment (CCSE),
Beihang University, Beijing, China
haogangzhu@buaa.edu.cn

[2] Key Laboratory of Data Science and Intelligent Computing, International
Innovation Institute, Beihang University, Hangzhou, China
zhangyingying@buaa.edu.cn

[3] Beijing Hospital/National Center of Gerontology of National Health Commission,
Beijing, 100730, China.
chuanyu2228@sina.com

[4] School of Instrumentation and Optoelectronic Engineering, Beihang University,
Beijing, 100191, China.

[5] School of Biological Sciences, Victoria University of Wellington, Wellington, New
Zealand.

Abstract. Echocardiography (ECHO) is commonly used to assist in the diagnosis of cardiovascular diseases (CVDs). However, manually conducting standardized ECHO view acquisitions by manipulating the probe demands significant experience and training for sonographers. In this work, we propose a visual navigation system for cardiac ultrasound view planning, designed to assist novice sonographers in accurately obtaining the required views for CVDs diagnosis. The system introduces a view-agnostic feature extractor to explore the spatial relationships between source frame views, learning the relative rotations among different frames for network regression, thereby facilitating transfer learning to improve the accuracy and robustness of identifying specific target planes. Additionally, we present a target consistency loss to ensure that frames within the same scan regress to the same target plane. The experimental results demonstrate that the average error in the apical four-chamber view (A4C) can be reduced to 7.055 degrees. Moreover, results from practical clinical validation indicate that, with the guidance of the visual navigation system, the average time for acquiring A4C view can be reduced by at least 3.86 times, which is instructive for the clinical practice of novice sonographers.

Keywords: Echocardiography · Visual navigation · Cardiac Ultrasound View Planning.

1 Introduction

Echocardiography is commonly used for diagnosing cardiovascular diseases (CVDs) owing to its real-time imaging, non-invasive nature, low cost, and convenience. [2, 4] However, the acquisition of standard views for CVDs diagnosis requires sonographers to establish the spatial correspondence between the dynamic three-dimensional structure of the heart and the two-dimensional ultrasound images within a limited examination time and acoustic window. This is a challenging task for novice sonographers. Therefore, automatic visual navigation for cardiac ultrasound view planning is high demand.

With the rapid development of deep learning, an abundance of studies on echocardiographic images have recently been conducted to optimize the scanning process, obtain the standard view and disease diagnosis. For instance, Narang et al. [9] utilized deep learning algorithm to assist nurses with no prior ultrasound experience in successfully capturing 10 echocardiographic views of diagnostic value. Wu et al. [14] introduced a knowledge distillation network to automatically and effectively identify 23 standard echocardiographic views commonly used in diagnosing congenital heart disease in children and achieved a good recognition effect. Grant et al. [3] employed spatiotemporal convolutions to conduct semantic segmentation of adult hearts and further categorized subtypes of CVDs. Hence, deep learning techniques have been proven to be highly effective in analyzing echocardiograms. However, the current analysis of cardiac ultrasound images fails to address the issue of visual navigation as it does not integrate positioning information. Acquiring positional information during cardiac ultrasound scanning faces two main challenges. First, significant individual differences make it difficult to establish a unified coordinate system for humans, leading to difficulties in obtaining positions. Second, because the heart is a dynamic organ, even minor variations in probe movement result in significant changes in the ultrasound imaging, making precise position capture difficult. This presents difficulties in achieving real-time synchronization between the cardiac ultrasound video and the data tracking probe movement.

In this work, we propose a human-based visual navigation system tailored for cardiac view planning. We first train a view-agnostic feature extractor to achieve features related to the three-dimensional structure of the heart from ultrasound frames and perform transfer learning for regression tasks across different views. Then, we introduce a new loss function to maintain consistency of the regression target views within the same video, where the target plane positions regressed from any two frames within the same scan should be identical. The main contribution of our work can be summarized as follows:

- We have designed a fully automated visual navigation system to guide probe movement for novice cardiac sonographers. To our knowledge, this is the first visual navigation for cardiac ultrasound view planning that can be used in real-world clinical scenarios on human body instead of phantom as used in previous studies.
- We proposed a view-agnostic feature extractor to explore the spatial relationships between abitrary frames from the same video, thereby enhancing

the accuracy and robustness of identifying various plane tasks. Furthermore, there is no need to retrain the feature extraction process from the beginning when introducing new planes that require navigation.

- We trained and validated our proposed framework with 3540 and 392 real human cardiac ultrasound scans, respectively. The experiments prove that it can effectively provide real clinical visual navigation for cardiac ultrasound view planning.

1.1 Related Works

Recently, there have been several attempts at visual navigation of echocardiography, which can be roughly divided into human-based and phantom-based methods. For human-based methods, Li et al. [6] introduced a reinforcement learning method for spine ultrasound standard view navigation, and achieved task success rates of 92% and 46% in intra-patient and inter-patient environments, respectively. Another work involved a dual-agent framework that combines reinforcement learning and deep learning to simulate the decision-making process of expert ultrasound physicians [7]. This framework autonomously acquires standard views during spinal ultrasound examinations and achieves an average navigation accuracy of 17.49° for inter-patient standard views in a simulated dataset collected from 17 volunteers. Yeung et al. [15] implemented a two-stage pre-training and fine-tuning network, which enabled a ultrasound navigation system based on fetal brain. The system achieved an effect where the euclidean distance could reach 23±9.01 voxels. For phantom-based methods, Zhao et al. [16] introduced a landmark retrieval-based ultrasound-probe movement guidance system utilizing data from the ScanTrainer Simulator to simulate the scanning process of obstetric ultrasound. Olivier et al. [10] utilized recurrent neural networks and visual attention to achieve probe movement guidance for acquiring standard views of apical four-chamber and parasternal long-axis in echocardiography. The error range for the x-axis lies between 4° and 15°, whereas for both the y-axis and z-axis, it is between 3° and 15°. All the above navigation methods have several limitations. Methods based on real human data are mostly focused on "static" organs such as the brain and spine. For dynamic organs like the heart, existing research relies on phantoms or simulators. However, the heart is a highly variable and dynamic organ with complex motion patterns that cannot be effectively simulated using static phantoms or simulators. It is difficult to replicate real movements and images, making clinical application challenging. To our knowledge, this is the first visual ultrasound navigation work on real human hearts.

2 Methods

2.1 Problem Setup

With a collection of N clinical scans, we obtain a series of the videos denoted as $\{S_v\}_{v=1}^{N}$. Each video, S_v, comprises L frames, along with their corresponding

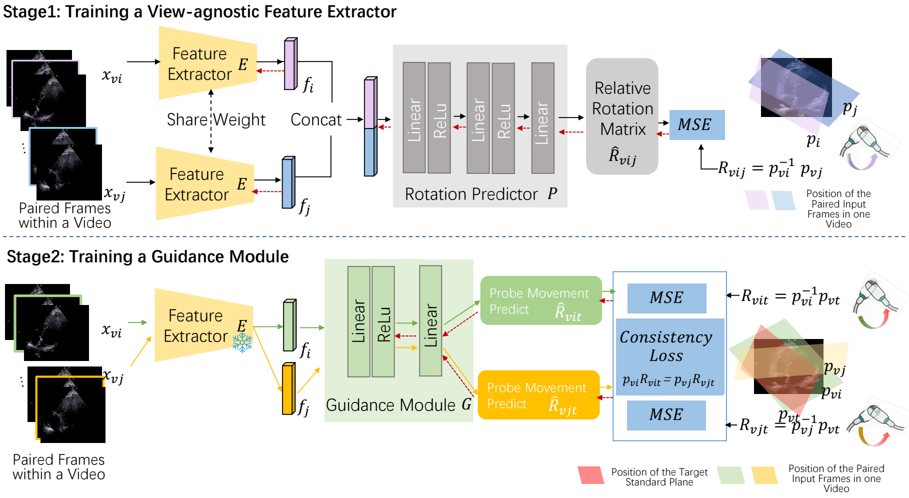
Stage1: Training a View-agnostic Feature Extractor
x_{vi}
Feature Extractor E
f_i
Share Weight
Concat
Feature Extractor E
f_j
x_{vj}
Paired Frames within a Video
Linear
ReLu
Linear
ReLu
Linear
Rotation Predictor P
Relative Rotation Matrix $\hat{R}_{vij}$
MSE
$R_{vij} = p_{vi}^{-1} p_{vj}$
p_j
p_i
Position of the Paired Input Frames in one Video
Stage2: Training a Guidance Module
x_{vi}
x_{vj}
Feature Extractor E
f_i
f_j
Linear
ReLu
Linear
Guidance Module G
Probe Movement Predict $\hat{R}_{vit}$
Probe Movement Predict $\hat{R}_{vjt}$
MSE
Consistency Loss
$p_{vi}R_{vit} = p_{vj}R_{vjt}$
MSE
$R_{vit} = p_{vi}^{-1}p_{vt}$
$R_{vjt} = p_{vj}^{-1}p_{vt}$
p_{vj}
p_{vi}
p_{vt}
Paired Frames within a Video
Position of the Target Standard Plane
Position of the Paired Input Frames in one Video

Fig. 1. Framework of the proposed network for real-world visual navigation. Stage 1: Train a view-agnostic feature extractor via regressing the rotation matrix of paired frames' positions within a video. Stage 2: Utilize the trained feature extractor to train the rotation matrix of the probe movement to the standard view via MSE loss and the proposed target consistency loss.

probe positions, represented by the set $\{(x_{vi}, p_{vi})\}_{i=1}^{L}$. Here, p_{vi} signifies the absolute rotation matrix relative to the standard position of the tracking system.

Visual navigation. Given an arbitrary input frame x_{vi}, the goal of visual navigation is to predict the probe rotation $\hat{R}_{vit}$ from the current probe position p_{vi} to the probe position p_{vt} of the anatomical standard plane, such as apical four-chamber view (A4C) and apical two-chamber view(A2C). The ground truth of the probe motion is calculated as:

$$R_{vit} = p_{vi}^{-1} p_{vt}, \tag{1}$$

where R_{vit} is the rotation matrix for the probe motion from the i_{th} frame to the target anatomical standard frame in video S_v.

To address the task of visual navigation, we propose a two-stage framework as illustrated in Figure 1. Initially, we train a view-agnostic feature extractor E and a rotation predictor P to compute the relative rotation matrix between probe positions of two frames in a single video. Subsequently, leveraging this learned feature extractor, we train a guidance module G to predict the probe's rotation relative to the standard plane.

2.2 View-agnostic Feature Extractor

We can directly regress the probe's rotation matrix towards the target anatomical standard plane to accomplish the visual navigation task. However, by directly conducting a regression model, the extracted features will be solely related to the target plane, ignoring the common features shared among different views. Re-training is required with every new target plane, wasting resources and making it less suitable for medical applications. Furthermore, the direct regression approach also fails to address the acquisition errors present in clinical scans collected by different physicians and the standard plane selections since the standard planes in ultrasound should be captured within a range of angles rather than a single specific position. Therefore, we propose a novel paradigm for navigation: initially pre-training a view-agnostic feature extractor that leverages intra-video relative probe position information, followed by fine-tuning a regression model to achieve the navigation. By utilizing the positional relationships between video frames, we not only extract more robust view-agnostic features but also enhance data efficiency. Our novel paradigm allows training on ultrasound data from any protocol, which is crucial for real-world navigation applications.

To efficiently extract view-agnostic features from paired input frames x_{vi} and x_{vj}, our approach utilizes a dual-backbone architecture where two networks operate in parallel, sharing the weights. These extracted features f_{vi} and f_{vj} are concatenated to form a combined representation of the image pair. Then, this concatenated feature is fed into a rotation predictor P, which predicts the relative rotation $\hat{R}_{vij}$ between the relative probe positions p_{vi} and p_{vj}. We use weighted least-squared error (MSE) as the loss function for the training:

$$\mathcal{L}_1 = \mathrm{MSE}(P(E(x_{vi}), E(x_{vj})), R_{vij}), \tag{2}$$

where $R_{vij} = p_{vi}^{-1} p_{vj}$. In detail, the rotation predictor P consists of two linear layers with 256 hidden units and a ReLU activation function and a final linear layer with 9 hidden units as the output layer. Subsequently, orthogonal Procrustes orthonormalization method, as detailed in RoMa [1], is applied to convert the 9-dimensional feature output into a standardized 3x3 rotation matrix.

2.3 Target View Relative Rotation Regression

Based on the pre-trained view-agnostic feature extractor E, we train the guidance module G according to the supervision of rotation matrices towards the target standard planes. The guidance module is designed with two fully connected layers separated by ReLU activations and the loss function is:

$$\mathcal{L}_2 = \mathrm{MSE}(G(E(x_{vi})), R_{vit}). \tag{3}$$

For the regression task, the final navigation target for each frame of the same clinical scan should be consistent: the position of the target plane. Therefore, we introduce a target consistency loss during the training process, defined as:

$$\mathcal{L}_{consist} = \mathrm{MSE}(p_{vi}\hat{R}_{vit}, p_{vj}\hat{R}_{vjt}), \tag{4}$$

note that p_{vi} and p_{vt} should belong to the same video. Ultimately, our navigation loss function combines the regression loss and the target consistency loss:

$$\mathcal{L}_{navi} = \mathcal{L}_2 + \lambda\mathcal{L}_{consist}, \tag{5}$$

where λ is the weight of the target consistency loss.

3 Experiments

3.1 Data acquisition

A total of 3932 clinical scans, including ultrasound videos and their corresponding probe motion trajectory data, were acquired at Anonymous Hospital. These scans were performed using a GE E95 scanner (*General Electric, USA*) equipped with an M5Sc-D probe, with the video frame rate set at 58 Hz. To track the probe's motion, we developed a motion-tracking system equipped with a VIPER4 positioning unit *(Polhemus, USA)*. The sensor of the positioning unit was attached to the probe with a 3D-printed mounting adapter. The probe orientation quaternions were sampled at 240Hz. The collected data is divided into two categories: A4C and A2C, totaling 2138 and 1794 clinical scans, respectively. For A4C view, scans begin at the standard A4C view, with the probe then rotating to a randomly generated point and subsequently returning to the standard A4C view. For A2C view, a similar protocol is followed. We selected 10% of the data as the validation set, resulting in 3540 training videos (1615 A4C, 1925 A2C) and 392 validation videos (179 A4C, 213 A2C). All video frames, originally 800x600, were resized to 224x224 for training. We collected a total of 741522 frames for A4C view and 353789 frames for A2C view. All the subjects used in this study are with ethical committee approval.

3.2 Experimental Settings

We adopt ResNet18 [5], ResNet50 [5] and MobileNetV2 [13] as the backbone networks for the view-agnostic feature extractor, respectively. We randomly select any two frames from a single scan as a pair for the network's input, and apply random Gaussian blur data augmentation to the input data. Gaussian kernel sizes are uniformly sampled from 5 to 9, and sigma ranges from 0.1 to 5.0. To balance the data from two target views, we replicated the frames related to A2C three times. The output dimension of the extracted image features from these networks is consistently set to 128. During the training process, the batch size is configured as 128, and each image is resized to 224x224 before input into the networks. For rotation-related computation, we use the RoMa [1] package. The AdamW [8] optimizer is used with a weight decay of 1×10^{-2}. The learning rate is initialized to 1×10^{-3} and gradually reduced to 1×10^{-4}. All the experiments are implemented with PyTorch [11] framework using Nvidia GPU V100. To evaluate the real-world performance of our visual navigation model, we developed a visual navigation testing system on the Nvidia Jetson AGX Orin developer kit. This system allows for the simultaneous visualization of real-time ultrasound scan images and provides guidance on probe positioning. It is built using the GTK4 GUI framework and the Nvidia Deepstream framework and then compiled and deployed using Nvidia TensorRT.

4 Results and Discussion

We use the geodesic distance d to evaluate the distance between the ground truth rotation R_{gt} and the predicted rotation $\hat{R}$ according to the previous work [12]. The geodesic distance d can be written as:

$$d(\hat{R}, R_{gt}) = cos^{-1}\left[\frac{tr(\hat{R}^{\top}R_{gt}) - 1}{2}\right]. \tag{6}$$

Table 1. Experimental results between different configurations and comparison results for view-agnostic feature extractor

Target View	Backbone	w/o VFE	w/ VFE
A4C	ResNet18	7.290	**7.139**
	ResNet50	7.389	**7.140**
	MobileNetV2	7.450	**7.208**
A2C	ResNet18	5.316	**5.203**
	ResNet50	6.045	**5.511**
	MobileNetV2	5.951	**5.329**

Table 2. Experimental results of target consistency loss

Target View	Backbone	w/o $L_{consist}$	w/ $L_{consist}$
A4C	MobileNetV2	7.208	**7.055**
A2C	MobileNetV2	5.329	**5.283**

Ablation Study. In this section, we evaluate the view-agnostic feature extractor and the target consistency loss function through ablation experiments. We first collected 200 series of standard view-random point-standard view acquisitions by cardiac ultrasound experts. The positions of the first and second standard views in the same acquisition were calculated, with an average rotation angle between the two views of 5.677 ± 4.117 degrees. As shown in Table 1, the view-agnostic feature extractor consistently enhances visual navigation accuracy across different backbone architectures, achieving a difference of only about 1 degree compared to expert cardiac sonographers. Specifically, when utilizing MobileNet V2 as the backbone, it yields improvements of 0.242 degree on the A4C view and 0.622 degree on the A2C view. We notice that the overall error in A2C is lower than in A4C. This could be due to the larger window required for obtaining the A4C view, which poses higher challenges in fitting due to the greater volatility in expert scanning. The results in Table 2 validate the effectiveness of the target consistency loss function, resulting in improvements of 0.153 degree on the A4C view and 0.046 degree on the A2C view with MobileNet V2.

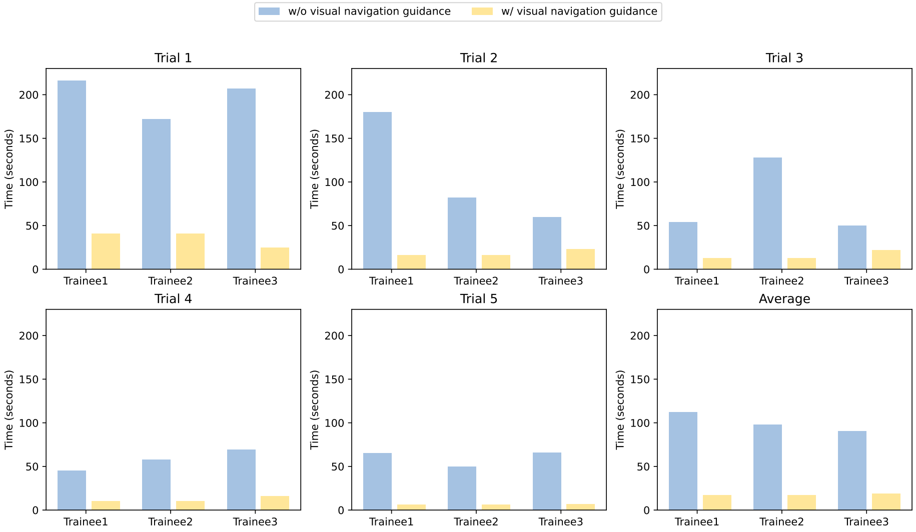

Fig. 2. Real-world scenario validation of A4C view planning with/without visual navigation guidance.

Real-world Verification for Visual Navigation. To validate the practical applicability of our proposed method, we selected three novice sonographers for real clinical validation. Before the actual clinical validation, a senior ultrasound physician first demonstrated and explained the fundamental knowledge and routine procedures for locating the standard A4C view. Each trainee was required to first locate the standard A4C plane without the aid of the visual navigation system, followed by locating the standard A4C view with the assistance of the visual navigation system. This sequence was repeated five times, with the duration of each attempt recorded. A one-minute break was allocated between each repetition, and any attempt exceeding three minutes was deemed a failure in plane localization. As shown in Figure 2, we can observe that with the guidance of the visual navigation system, the average time taken for each of the three trainees' five operations was reduced by at least 3.86 times. Furthermore, while each trainee experienced one failed plane localization out of five attempts without the visual navigation system, all five plane localizations were successful with the assistance of the visual navigation system (see the supplementary materials for details).

5 Conclusion

In this paper, we present the first real-world visual navigation system based on the real human heart. This system consists of two parts: a view-agnostic feature extractor and a regression module. The view-agnostic feature extractor learns the relative spatial relationships between ultrasound video frames and then transfers this knowledge to the regression module to enhance the regression accuracy across different views. In the specific-view regression tasks, we introduce a target consistency loss to maintain the consistency of the target views within a single scan, thus further improving the accuracy of visual navigation. Experimental results suggest that our proposed method can consistently improve the accuracy of cardiac ultrasound visual navigation. Furthermore, in practical clinical validation, our visual navigation system is shown to assist novice sonographers in accurately locating the A4C view and reducing the time required for view localization by 3.86 times. Therefore, our approach is crucial for clinical applications in visual navigation for cardiac ultrasound view planning, and in the future, we can extend it to other views of the heart as well as other organs.

Acknowledgments. This work is supported in part by the Major Project of Science and Technology Innovation 2030 - New Generation Artificial Intelligence under Grant 2021ZD0140407, in part by the National Natural Science Foundation of China under Grant U21A20523, in part by the Beijing Natural Science Foundation under Grant (L222152, 7244325).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Brégier, R.: Deep regression on manifolds: a 3d rotation case study. In: 2021 International Conference on 3D Vision (3DV). pp. 166–174. IEEE (2021)
2. Dave, J.K., Mc Donald, M.E., Mehrotra, P., Kohut, A.R., Eisenbrey, J.R., Forsberg, F.: Recent technological advancements in cardiac ultrasound imaging. Ultrasonics **84**, 329–340 (2018)
3. Duffy, G., Cheng, P.P., Yuan, N., He, B., Kwan, A.C., Shun-Shin, M.J., Alexander, K.M., Ebinger, J., Lungren, M.P., Rader, F., et al.: High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. JAMA cardiology **7**(4), 386–395 (2022)
4. Feigenbaum, H.: Evolution of echocardiography. Circulation **93**(7), 1321–1327 (1996)
5. 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)
6. Li, K., Wang, J., Xu, Y., Qin, H., Liu, D., Liu, L., Meng, M.Q.H.: Autonomous navigation of an ultrasound probe towards standard scan planes with deep reinforcement learning. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 8302–8308. IEEE (2021)
7. Li, K., Xu, Y., Wang, J., Ni, D., Liu, L., Meng, M.Q.H.: Image-guided navigation of a robotic ultrasound probe for autonomous spinal sonography using a shadow-aware dual-agent framework. IEEE Transactions on Medical Robotics and Bionics **4**(1), 130–144 (2021)
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

9. Narang, A., Bae, R., Hong, H., Thomas, Y., Surette, S., Cadieu, C., Chaudhry, A., Martin, R.P., McCarthy, P.M., Rubenson, D.S., et al.: Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. JAMA cardiology **6**(6), 624–632 (2021)
10. Olivier, D., McGuffin, M.J., Laporte, C.: Utilizing sonographer visual attention for probe movement guidance in cardiac point of care ultrasound. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
12. Salehi, S.S.M., Khan, S., Erdogmus, D., Gholipour, A.: Real-time deep pose estimation with geodesic loss for image-to-template rigid registration. IEEE transactions on medical imaging **38**(2), 470–481 (2018)
13. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
14. Wu, L., Dong, B., Liu, X., Hong, W., Chen, L., Gao, K., Sheng, Q., Yu, Y., Zhao, L., Zhang, Y.: Standard echocardiographic view recognition in diagnosis of congenital heart defects in children using deep learning based on knowledge distillation. Frontiers in Pediatrics **9**, 770182 (2022)
15. Yeung, P.H., Aliasi, M., Haak, M., 21st Consortium, I., Xie, W., Namburete, A.I.: Adaptive 3d localization of 2d freehand ultrasound brain images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 207–217. Springer (2022)
16. Zhao, C., Droste, R., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Visual-assisted probe movement guidance for obstetric ultrasound scanning using landmark retrieval. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. pp. 670–679. Springer (2021)