Contents lists available at ScienceDirect

Medical Image Analysis

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

A Bayesian network for simultaneous keyframe and landmark detection in ultrasonic cine

Yong Feng [a,b], Jinzhu Yang [a,b,c,*], Meng Li [d], Lingzhi Tang [a,b], Song Sun [a,b], Yonghuai Wang [d]

[a] *Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, Shenyang, China*
[b] *School of Computer Science and Engineering, Northeastern University, Shenyang, China*
[c] *National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Shenyang, China*
[d] *Department of Cardiovascular Ultrasound, The First Hospital of China Medical University, Shenyang, China*

ARTICLE INFO

Keywords:
Landmark detection
Keyframe detection
Bayesian network
Ultrasonic cine
Multi-task learning

ABSTRACT

Accurate landmark detection in medical imaging is essential for quantifying various anatomical structures and assisting in diagnosis and treatment planning. In ultrasound cine, landmark detection is often associated with identifying keyframes, which represent the occurrence of specific events, such as measuring target dimensions at specific temporal phases. Existing methods predominantly treat landmark and keyframe detection as separate tasks without harnessing their underlying correlations. Additionally, owing to the intrinsic characteristics of ultrasound imaging, both tasks are constrained by inter-observer variability, leading to potentially higher levels of uncertainty. In this paper, we propose a Bayesian network to achieve simultaneous keyframe and landmark detection in ultrasonic cine, especially under highly sparse training data conditions. We follow a coarse-to-fine landmark detection architecture and propose an adaptive Bayesian hypergraph for coordinate refinement on the results of heatmap-based regression. In addition, we propose Order Loss for training bi-directional Gated Recurrent Unit to identify keyframes based on the relative likelihoods within the sequence. Furthermore, to exploit the underlying correlation between the two tasks, we use a shared encoder to extract features for both tasks and enhance the detection accuracy through the interaction of temporal and motion information. Experiments on two in-house datasets (multi-view transesophageal and transthoracic echocardiography) and one public dataset (transthoracic echocardiography) demonstrate that our method outperforms state-of-the-art approaches. The mean absolute errors for dimension measurements of the left atrial appendage, aortic annulus, and left ventricle are 2.40 mm, 0.83 mm, and 1.63 mm, respectively. The source code is available at github.com/warmestwind/ABHG.

1. Introduction

Ultrasonic cine, a common form of preservation in ultrasonography, allows real-time review of moving structures and flow patterns, providing additional diagnostic information and enabling subtle abnormality identification (Mitchell et al., 2019). In a clinical routine, manually identifying the event of interest within an ultrasonic cine and performing further measurements and diagnoses are fundamental tasks (Lang et al., 2015). However, performing this repetitive task can be challenging, even for experienced professionals, especially when dealing with intricate anatomical variations, sub-optimal image quality, or high caseloads of patients. As a result, there is an increasing demand for automated assistance tools that can streamline and enhance the accuracy of this critical process (Ouyang et al., 2020).

Automatic identification of cardiac cycle phases in echocardiography for measuring various cardiac parameters is a significant clinical application that can reduce inter-observer variability, especially in point-of-care scenarios where electrocardiographic data might be absent (Ciusdel et al., 2020; Zhang et al., 2023). Dezaki et al. (2017) propose structured loss based on the changes in heart volume to identify end-diastolic (ED) and end-systolic (ES) frames. Dai et al. (2023) exploit the cyclical nature of heart pumping to design self-supervised regularization terms such that the features of frames exhibit temporal cyclicality. However, these methods achieve keyframe detection, assuming cardiac volume varies monotonically within different temporal phases. We believe that models relying on this assumption might lack adaptability for patients with arrhythmia, necessitating the design of more generalized keyframe detection models. In addition, for quantitative description tasks, a frame with clear target morphology has to be found in the ultrasonic cine. Therefore, more target features, such

[*] Corresponding author at: Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, Shenyang, China.
E-mail addresses: yangjinzhu@cse.neu.edu.cn (J. Yang), wyh_wyh2012@163.com (Y. Wang).

https://doi.org/10.1016/j.media.2024.103228
Received 30 October 2023; Received in revised form 28 May 2024; Accepted 30 May 2024
Available online 1 June 2024

as texture, temporal, or spatial features, need to be introduced (Wang et al., 2022).

After identifying the keyframe, a landmark detection model is introduced to locate critical anatomical structures (Mokhtari et al., 2023). Xu et al. (2018) utilize an encoder–decoder to construct a multi-task network for simultaneous ultrasound abdominal view classification and heatmap-based landmark detection. In order to overcome the effects of irrelevant noise and echo attenuation, Xu et al. (2021) realize the infant hip landmark detection by calculating the relation matrix and mining the long-range dependency on the basis of heatmap regression. For the same purpose, the attention mechanism is applied to landmark detection at muscle–tendon junction (Leitner et al., 2022) and pleura (Tripathi et al., 2023) in ultrasonic cine. However, these methods operate independently from the keyframe detection model and do not leverage the potential correlation between the two tasks. This separation may result in suboptimal performance or limited adaptability.

Keyframe and landmark detection tasks within ultrasound cine exhibit inherent potential for substantial uncertainty (Jafari et al., 2022). Keyframes, capturing pivotal moments, stand susceptible to temporal variations and intricate anatomical complexities, often entailing a degree of subjectivity in their selection. Simultaneously, landmark detection grapples with the challenges posed by noise, artifacts, and distortions within the image. Moreover, the compounding influence of inter-observer variability and constrained resolution further exacerbates the overarching uncertainty inherent in the process. Recent studies show that using uncertainty quantification methods, such as Monte Carlo Dropout (MC Dropout) or Bayesian methods, can provide a more nuanced understanding of the uncertainty and yield more consistent and reliable predictions (Zhang et al., 2019).

In this work, we propose a unified Bayesian network to detect keyframes and landmarks within ultrasonic cine simultaneously. We argue that the accurate identification of keyframes in dynamic sequences should be referenced to the spatial and temporal attributes of landmarks, while the precise localization of landmarks can be enhanced through temporal constraints derived from keyframes. To achieve precise landmark localization, we adopt a coarse-to-fine framework. Initially, a shared encoder–decoder is employed to extract frame features and perform coarse localization through heatmap regression. Subsequently, we propose an adaptive Bayesian hypergraph to perform coordinate fine-tuning based on the coarse localization. Moreover, when selecting the keyframe, we integrate motion information of landmarks within the dynamic sequence with image features. Notably, in contrast to prior methods assigning specific values for frame indexing, we solely establish the relative order within the sequence. Accordingly, we introduce Order Loss for bi-directional Gated Recurrent Unit (Bi-GRU) (Chung et al., 2014) training to identify keyframe.

By leveraging the inherent correlations between the two tasks, we alleviate the high uncertainty that arises from a combination of factors intrinsic to the imaging modality and improve accuracy. The main contributions of this paper are summarized as follows:

1. We propose a multi-task framework to simultaneously detect keyframes and landmarks in ultrasonic cine, where the keyframes detection integrates the motion state of landmarks, while the landmarks detection is constrained by temporal characteristics.
2. To improve the performance of single-stage landmark detection models, we propose an adaptive Bayesian hypergraph following a coarse-to-fine architecture.
3. For the keyframe identification task, we design Order Loss, which consists of a partial log-likelihood loss to limit the relative magnitude within a sequence, and a triplet loss as a regularization term to distinguish the difference between keyframes and non-keyframes.
4. We conduct experiments on both transesophageal and transthoracic echocardiographic datasets for multi-view left atrial appendage, aortic annulus, and left ventricle anatomical landmarks and keyframe detection. Our proposed method outperforms state-of-the-art methods on both tasks, demonstrating its great potential for clinical applications.

2. Related work

2.1. Keyframe detection

The keyframes in an ultrasonic cine capture significant events, such as the appearance of target lesions in diagnostic imaging. Considering the temporal characteristics of ultrasonic cine, Wang et al. (2022) take three types of features (texture, position, and temporal index) of thyroid nodules into a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) to predict keyframes supporting the malignancy diagnosis and quantitative analysis. Huang et al. (2022) introduce reinforcement learning to the keyframe detection task, training an agent to determine whether ultrasound frames are selected as keyframes. The proposed model leverages tumor detection confidence and malignant signs to further enhance subsequent diagnostic accuracy.

In addition, identifying the representation of image features under different cardiac motion cycles is crucial in cardiovascular ultrasound examinations (Lane et al., 2021; Pu et al., 2021; Jahren et al., 2020). Wang et al. (2021a) propose a multi-task framework for evaluating right ventricular function through keyframe and landmark recognition. This framework employs a shared encoder to extract features from ultrasonic cine and utilizes a decoder for right ventricular landmark localization. The encoder features are also processed through global average pooling and fed into a Recurrent Neural Network (Elman, 1990) to regress scores for each frame related to the ES and ED in the echocardiography sequence. Meanwhile, recognizing cardiac phases requires considering model structure, label definition, and loss function design. Kong et al. (2016) introduce temporal structured loss to constrain the monotonicity of consecutive frames. Furthermore, T. Dezaki et al. (2019) emphasize overall consistency, setting the ED and ES frames as the global maximum and minimum values, respectively. Additionally, Ciusdel et al. (2020) consider frame index regression as a keyframe classification task, but an important aspect to take into account is the imbalance between keyframes and non-keyframes, especially in small-scale datasets.

2.2. Landmark detection

Traditional landmark detection methods hinge on meticulously crafted feature descriptors (Maraci et al., 2017) and well-defined detectors (Sedai et al., 2015) to analyze unique image features. For their simplicity and efficiency, some statistical models (Rueda and Alcañiz, 2006) and template-matching methods (Quan et al., 2022) remain applicable for anatomical landmark localization. Recently, deep learning-based landmark detection methods have achieved remarkable success in face recognition (Zou et al., 2019), pose recognition (Xu et al., 2022), cephalometric analysis (Oh et al., 2021), brain function analysis (He et al., 2023), and automated measurement of clinical parameters (Xu et al., 2021). These localization methods can be broadly categorized into direct regression-based and heatmap-based methods.

The regression-based approach directly predicts coordinates or coordinate offsets by extracting visual features (McCouat and Voiculescu, 2022). DeepPose (Toshev and Szegedy, 2014), a classical coordinate regression model, extracts image features with AlexNet (Krizhevsky et al., 2012) and directly regresses landmark coordinates. Noothout et al. (2020) achieve coordinate refinement by predicting the offset displacement from the center point of an image patch to the target point, along with the importance of that patch. Alansary et al.

(2019) rely on a reinforcement learning strategy to train multiple intelligences to identify the optimal path to reach a landmark. On the other hand, the heatmap-based approach can be seen as a semantic segmentation of landmarks, where the semantic mask is a Gaussian distribution centered on each landmark (Gilbert et al., 2019). A classical heatmap-based human landmark detection model is based on a stacked hourglass structure (Newell et al., 2016) with constant bottom-up and top-down processing to integrate multi-scale features for better spatial information capture. Considering that the heatmap-based approach eventually determines coordinates in an un-derivable way, i.e., maximizing the likelihood, Sun et al. (2018) propose a derivable heatmap-based regression method, which further narrows the gap between the heatmap-based and regression-based methods.

In recent years, a coarse-to-fine landmark detection architecture has been further explored. Li et al. (2020) propose a two-stage adaptive graph model, where the first stage learns a perspective transformation matrix with the mean value of landmarks in the training set as the initial value; the second stage uses the same graph structure to fine-tune the coarse localization results of the first stage. Lang et al. (2020) also adopt a coarse-to-fine architecture, where an encoder–decoder structure is used to obtain landmark heatmaps in the coarse localization phase. Then, the obtained heatmaps are weighted with the decoded features as inputs to a graph model to obtain more accurate 3D coordinates. Further, Chen et al. (2022) use LSTM in the fine stage to iteratively update the local and global features of each landmark to fine-tune the heatmap-based regression results. Lu et al. (2022) use multi-scale features and spatial information to train a graph model with a learnable adjacency matrix and iteratively obtain an accurate prediction. In response to the domain shift issue, Jin et al. (2023) propose an unsupervised domain-adaptive anatomical landmark detection framework, which incorporates transformer-based coarse localization and heatmap-based local refinement.

2.3. Bayesian model and uncertainty

Bayesian models offer a principled and rigorous framework for addressing various challenges in machine learning. They provide robustness to noisy data, mitigate overfitting, and effectively control model complexity by incorporating prior knowledge. Additionally, Bayesian methods are crucial for estimating model uncertainty in domains like assisted diagnosis (Eaton-Rosen et al., 2018) and autonomous driving (Kendall et al., 2017), where comprehending uncertainty is paramount.

While Bayesian inference accurately assesses evidence, it is constrained by computational complexity, posing challenges for its application in deep models. Several computational approximations have been proposed for Bayesian modeling and uncertainty estimation (Jospin et al., 2022). Markov Chain Monte Carlo (Geyer, 1992), a sampling-based approach, provides approximate samples from the posterior distribution. Meanwhile, Variational Inference, an optimization-based approach, aims to find an approximate simple distribution that closely matches the true posterior (Graves, 2011); MC Dropout (Gal and Ghahramani, 2016), an approximation to the probabilistic deep Gaussian process, which utilizes dropout during test-time (MC Dropout) to estimate model uncertainty through multiple forward passes. Additionally, Deep Ensembles (Lakshminarayanan et al., 2017) utilizes an ensemble-based strategy by training multiple instances of the same model architecture with varied initializations or training data subsets to achieve reliable uncertainty estimates. These approaches collectively contribute to advancing our understanding of model uncertainty in deep learning applications.

Bayesian modeling has garnered significant attention in the field of medical imaging. Jafari et al. (2021) propose a Bayesian U-Net approach for left ventricle segmentation in ultrasound cine by employing MC Dropout to generate an uncertainty ring around the segmentation boundary. They further observe higher Epistemic and Aleatoric uncertainty in deep Bayesian models for landmark detection in non-keyframe predictions, particularly in sequences with significant variations in target organ morphology (Jafari et al., 2022). Schobs et al. (2023) introduce a clinical application quality control method called Quantile Binning for heatmap-based landmark detection models. Hiasa et al. (2020) apply the Bayesian U-Net for lower limb muscle segmentation, establishing the correlation between uncertainty and segmentation failures. While previous research has highlighted the efficacy of Bayesian modeling for static data, more comprehensive studies on sparsely labeled dynamic data in this domain still need to be completed.

3. Method

3.1. Problem definition

We aim to identify the keyframe in the ultrasonic cine and locate landmarks in the keyframe to measure anatomical parameters ultimately. For this multi-task model M, we represent the training data $D_{train} = \{(X_i, I_i, Y_i)\}_{i=1}^{N}$, where $X_i = \{x_1 \dots x_T\}$ denotes the ith cine sample, T is the number of frames, I_i denotes the index of keyframe, and Y_i denotes the landmark coordinates in the keyframe. We remark that the annotation of each training sample is sparse, containing only one keyframe and a set of landmark annotations on it. In the test phase of this sparse problem, for an ultrasonic cine X_t, the model predicts its keyframe and landmarks, i.e., $(\hat{I}_t, \hat{Y}_t) = M(X_t)$. The overview of proposed multi-task detection model is shown in Fig. 1 and further detailed below.

3.2. Landmark detection

We adopt a coarse-to-fine architecture to detect landmarks. First, we extract the image features of each frame through an encoder–decoder and perform a heatmap-based coarse localization. Then, we take this coarse result as the initial point and introduce a hypergraph to model the structural relationship of landmarks for further optimization explicitly.

3.2.1. Heatmap-based coarse localization

In the coarse localization stage, we directly apply a three-by-three convolution and a softmax to the global features $F_{global} \in R^{C\times H\times W}$ extracted by the decoder to get the heatmap $Z_k \in R^{H\times W}$ of the k th landmark. Then, the coarse prediction coordinates L_k are obtained by heatmap-based regression (Sun et al., 2018):

$$L_k = \int_P P \cdot Z_k(P), \tag{1}$$

where P denotes the pixel coordinates in the heatmap. Although the encoder–decoder has been widely used in segmentation, the global features have limited ability to estimate landmark coordinates and the relationships between them. Therefore, we introduce a graph model to optimize the predictions further.

3.2.2. Graph convolution process

Considering a graph $G = \{A, F\}$, where the adjacency matrix $A = \{a_{ij}\} \in R^{n\times n}$ describes the pair-wise relations of n nodes, and $F \in R^{n\times d}$ denotes the d dimensional features of each node. The conventional graph convolution (Kipf and Welling, 2017) for l level nodes can be expressed as neighborhood information aggregation along with linear W and nonlinear σ transformations, i.e., $F^{l+1} = \sigma(AF^l W^{l+1})$.

Hypergraph extends traditional graph structure by incorporating hyperedges (Zhou et al., 2006), providing a more flexible and expressive framework for representing node relationships. A hypergraph can be represented as $HG = \{H, \Lambda, F\}$, where H is a $n \times ne$ binary incidence matrix and ne denotes the number of hyperedge. The entry $H(v, e)$ is equal to 1 if hyperedge e is connected to node v and 0 otherwise. The

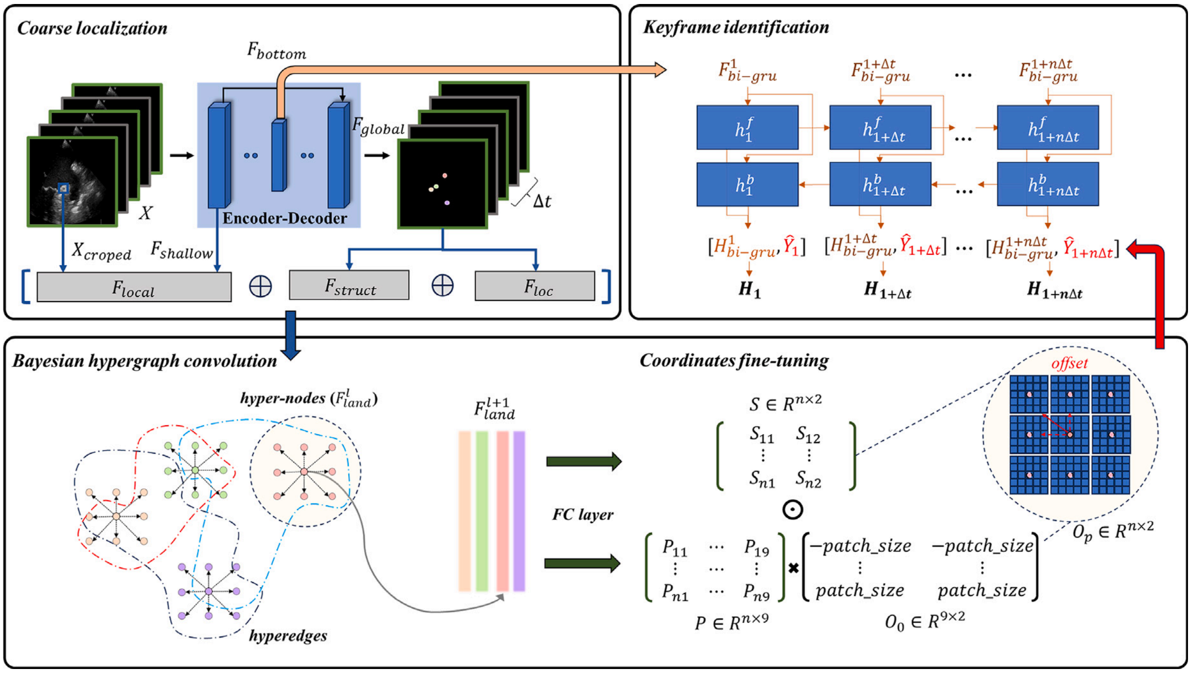

Fig. 1. The proposed architecture simultaneously identifies landmarks and keyframe for ultrasonic cine. First, a shared encoder–decoder is utilized to extract the features of each frame and perform coarse localization of landmarks. Next, the Bayesian hypergraph is adaptively constructed by utilizing the local information of the original image, the shallow features of the encoder, the structural relationships between landmarks and the coarse locations as the node features. To achieve accurate coordinate fine-tuning, we use Bayesian hypergraph convolution to establish hypernode connections within 8-neighborhoods and predict the direction and scale of the offset. Finally, the predicted landmark coordinates and coder–decoder global features are fed into a Bi-GRU to obtain the likelihoods of keyframes.

diagonal matrix $\Lambda = diag(w_1, \dots, w_{ne})$ defines the hyperedge weights. Then, the hypergraph convolution (Feng et al., 2019) can be expressed as: $F^{l+1} = \sigma(D_v^{-1/2} H \Lambda D_e^{-1} H^T D_v^{-1/2} F^l W^{l+1})$, where D_v and D_e denote the diagonal matrices of the vertex degrees and hyperedge degrees respectively.

However, considering the constrained size of the graph composed of landmarks in medical images, which leads to limited information transfer and is prone to overfitting, we construct an Adaptive Bayesian Hypergraph (ABHG) to model the relations among landmarks more flexibly and effectively.

3.2.3. Fine-tuning with ABHG

We define $ABHG = \{T, \tilde{H}, \Lambda, F_{land}\}$. It can be noticed that we add an aggregation matrix $T \in R^{n\times nh}$ for the adaptively expanded hyper-nodes. As shown in Fig. 2, we incorporate nodes with a predefined offset $O_0 \in R^{(8+1)\times 2}$ in the 8-neighborhood direction of each landmark, adaptively scaling the hyper-node size from n to $n(8+1)$. Each hyper-node feature F_{land} consists of a local feature $F_{local} \in R^c$, a structural feature $F_{struct} \in R^{n(8+1)}$, and a location feature $F_{loc} \in R^2$, i.e., $F_{land} = F_{local} \oplus F_{struct} \oplus F_{loc}$. Specifically, the local features F_{local} are obtained by cropping and convolving the 8 × 8 sized patches centered on the predicted locations from the previous stage on the original image and the shallow features $F_{shallow}$ of the encoder; The structural feature F_{struct} is the predicted coordinate differences between landmarks; The location feature F_{loc} are predicted coordinates. Further, our hypergraph convolution can be expressed as:

$$F^{l+1}_{land} = \sigma(\gamma^l T \tilde{H} \Lambda \tilde{H}^T F^l_{land} W^{l+1}), \tag{2}$$

where $\tilde{H}$ is a learnable incidence matrix, γ^l is a vector of independent Bernoulli random variables each of which has probability p of being 1. We leverage the MC Dropout mechanism at test time to estimate model uncertainty by running multiple stochastic forward passes.

To achieve fine-tuning of point coordinates, we utilize two fully connected layers to the hyper-nodes for predicting the shifting probabilities $P \in R^{n\times(8+1)}$ of landmarks at 8 + 1 offset directions and the corresponding scales $S \in R^{n\times 2}$, respectively. Then, the predicted offset O_p for the landmarks fine-tuning can be expressed as:

$$O_p = PO_0 \odot S. \tag{3}$$

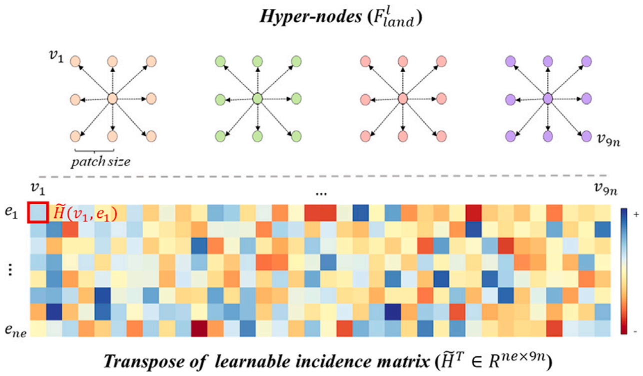
Hyper-nodes (F^l_{land})
v_1
patch size
v_{9n}
v_1
...
v_{9n}
e_1
$\tilde{H}(v_1, e_1)$
⋮
e_{ne}
+
-
Transpose of learnable incidence matrix ($\tilde{H}^T \in R^{ne\times 9n}$)

Fig. 2. The 8-neighborhood hyper-nodes and the incidence matrix of ABHG.

3.3. Keyframe detection

For the keyframe detection task, in addition to model design, we have to define the ground truth and design a plausible training process. In this section, we first define a temporal label based on the relative relationships within a cine. Then, we model the temporal relations with a Bi-GRU and predict the keyframe with the temporal variations of the landmarks. Finally, we design a loss function that conforms to the temporal characteristics of ultrasonic cine.

3.3.1. Ground-truth definition

For descriptive convenience, we define each ultrasonic cine X_i containing only one keyframe X_i^K, where $k \in [1, T]$. In particular, we intercept one of the cycles for a sequence with a periodicity like echocardiography. It is important to note that our approach can be extended to multi-cycle sequences and multi-keyframe scenarios. Unlike the conventional method of assigning fixed indices to frames, we define keyframe labels based on their relative order within the sequence. Specifically, we assign labels $I_i^1 > \cdots > I_i^K$ and $I_i^T > \cdots > I_i^K$, where I_i^K represents the keyframe with the minimum label value. The label values increase as frames move further away from the keyframe. We use a parabolic tendency to define index labels, allowing us to capture the relative importance of frames in a sequence and thus provide a more nuanced representation of the keyframe. By considering relative order rather than absolute position, we accommodate variability across sequences and maintain consistency within each ultrasound sequence.

3.3.2. Bi-GRU

In the context of identifying keyframes in ultrasonic cine, we employ a sequence model, Bi-GRU, which has been shown to be advanced and widely adopted (T. Dezaki et al., 2019). As an improved recurrent neural network, Bi-GRU can analyze the input sequence from forward and reverse. This bidirectional processing is particularly valuable for keyframe detection, as it allows for a more comprehensive understanding of the video sequence, potentially enhancing the accuracy of keyframe identification. The bidirectional processing of Bi-GRU is defined as follows:

$$\begin{aligned} h_t^f &= GRU_f(X_i, h_{t-1}^f) \\ h_t^b &= GRU_b(\overleftarrow{X}_i, h_{t-1}^b) \\ H_{bi-gru}^t &= h_t^f + h_t^b, \end{aligned} \tag{4}$$

where $\overleftarrow{X}$ represents the video sequence in reverse order. The t state of the Bi-GRU H_{bi-gru}^t is obtained by adding the states of the forward and the backward GRU (Cho et al., 2014). Eventually, the activation value H_t for each frame in the keyframe branch is obtained by applying a 1 × 1 convolution to the concatenated features of H_{bi-gru}^t and the predicted landmark coordinates $\hat{Y}_t$, followed by a *tanh* activation function.

3.3.3. Loss function

The previous training methods for keyframe detection can be broadly categorized into two main approaches. One approach treats keyframe detection as a binary classification problem, where the model is trained to classify each frame as a keyframe or non-keyframe using cross-entropy loss. This approach leverages the power of classification algorithms to learn discriminative features and make accurate predictions. On the other hand, some methods consider keyframe detection as a regression task, assigning a real-valued score to each frame to indicate its likelihood of being a keyframe based on relevance or importance. These models are trained using mean squared loss, allowing for a fine-grained estimation of frame importance. The regression approach offers more flexibility in capturing the relative significance of frames.

We introduce the partial log-likelihood loss (PL Loss) as the primary loss, often used in survival analysis (Tang et al., 2024), to train our keyframe detection branch. In survival analysis, we often encounter censored data, where the event of interest (e.g., death) has not occurred by the end of the study or observation period. The partial likelihood compares the relative order of event times rather than the exact timing. It assigns a higher likelihood to the model that predicts the observed events in the correct order while accommodating censored data. By using the partial log-likelihood loss, we aim to effectively capture the temporal patterns and relevance of frames in the context of keyframe detection. The partial log-likelihood loss is defined as follows:

$$\begin{aligned} log(PL(H_t)) &= log(\prod_{j=1}^{K} \frac{e^{H_j}}{\sum_{i=1}^{j} e^{H_i}}) + log(\prod_{m=T}^{K} \frac{e^{H_m}}{\sum_{n=T}^{m} e^{H_n}}) \\ &= \sum_{j=1}^{K} H_j - log(\prod_{i=1}^{j} e^{H_i}) + \sum_{m=T}^{K} H_m - log(\prod_{n=1}^{m} e^{H_n}), \end{aligned} \tag{5}$$

where e represents the natural constant and H_t denotes the logit output of the keyframe branch for the t th frame.

We also adopt the triplet loss (Schroff et al., 2015) as a regularization term to encourage the model to learn discriminative features that separate keyframes from non-keyframes. We randomly sample two frames from the input sequence and define the one with a shorter temporal distance to the keyframe a as the positive sample p, while the one with a greater distance as the negative sample n. Naturally, the keyframe is used as the anchor sample. Then, the triplet loss is defined as follows:

$$\begin{aligned} TriL(a,p,n) = max(&\|F_{bottom}^{p} - F_{bottom}^{a}\|_2^2 \\ &- \|F_{bottom}^{n} - F_{bottom}^{a}\|_2^2, 0), \end{aligned} \tag{6}$$

where F_{bottom} represents the bottom feature of the encoder–decoder, as shown in Fig. 1.

Eventually, we train the keyframe detection with Order Loss, which comprises the partial log-likelihood loss and the randomized triplet regularization:

$$\mathcal{L}_{Order} = log(PL(H_t)) + \alpha TriL(a,p,n). \tag{7}$$

In this study, we set α to 0.1 by cross-validation to balance two competing objectives.

4. Experiments and evaluations

4.1. Datasets

We validate the proposed multi-task model on two sparsely annotated ultrasonic cine datasets, LAA and PLAX, as shown in Fig. 3. The LAA dataset comprises 261 B-mode ultrasonic cines of the left atrial appendage captured from transesophageal echocardiography scans, providing different perspectives including 0°, 45°, and 90° (with 59, 103, and 99 instances, respectively). We capture a complete cardiac cycle (average 53 frames) from each cine and select the left ventricular end-systolic phase as the keyframe (average frame index of 26.2 ± 7.7) to annotate the four landmarks for measuring orifice diameter (LAAO) and depth (LAAD) of the left atrial appendage. It is worth noting that the LAA dataset is collected from patients with atrial fibrillation, comprising 203 sequences from paroxysmal patients and 58 from those with persistent atrial fibrillation. These sequences allow for a comprehensive analysis of the morphology and function of the LAA. The PLAX dataset includes 153 parasternal long-axis view cines captured from B-mode transthoracic echocardiography scans. We capture a complete cardiac cycle (average 54 frames) from each cine and select the left ventricular mid-systolic phase as the keyframe (average frame index of 7.7 ± 3.4) to annotate the two landmarks of the aortic annulus (AA). PLAX provides an excellent view of the aortic valve, allowing assessment of the valve leaflets and annulus. All data are double-blind annotated by two junior physicians and confirmed by one senior physician. We believe that these two datasets are valuable resources for evaluating and benchmarking our proposed methods for keyframe and landmark detection tasks of ultrasonic cine.

Additionally, a larger public ultrasound dataset named UIC is employed to further assess the accuracy of landmark localization (Howard et al., 2021). UIC comprises diverse echocardiographs collected from 17 local UK hospitals, collaboratively annotated by echocardiographers and cardiology experts to measure parameters such as the interventricular septum (IVS), left ventricular internal diameter (LVID), and posterior wall (PW). On the parasternal long-axis view of UIC, four landmarks are annotated (top and bottom of the IVS, top and bottom of the PW), as shown in Fig. 3. We follow the original data partitioning, excluding samples with incomplete coordinates in the training and validation sets, as well as those lacking essential physical information in the test set, to ensure the accuracy of both training and testing. The filtered training, validation, and test sets contain 1438, 337, and 249 images. It is important to note that the UIC dataset only includes single-frame images rather than ultrasonic cines.

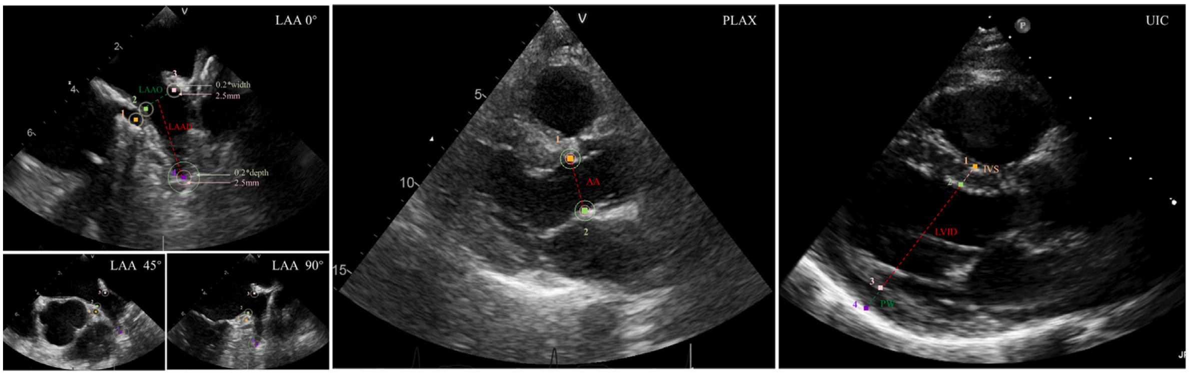

Fig. 3. In the two in-house datasets, LAA and PLAX, we annotate four anatomical landmarks on the left atrial appendage of multi-view transesophageal echocardiography: (1) the left circumflex branch of the coronary artery, (2) the inner wall of the left auricular wall, (3) a landmark located 2 mm below the opening of the left upper pulmonary vein on the opposite side, and (4) the tip of the left atrial appendage. Additionally, two anatomical landmarks, namely (1) the anterior inner edge and (2) the posterior inner edge, were identified in the parasternal long-axis view of transthoracic echocardiography. In the public dataset UIC, four landmarks are annotated: the top and bottom of the interventricular septum, as well as the top and bottom of the posterior wall.

4.2. Implementation details

Considering the scale of the dataset and the sparsity of the annotation, we perform a five-fold cross-validation on both datasets. Each frame of the ultrasonic cine is normalized to 0-1 and adjusted to 256 × 256 pixels for a uniform experimental setup. Data augmentation methods include horizontal flip, Gaussian blur, translation, rotation, scaling, and gamma transformation. Our model is implemented on a NVIDIA A100 using Pytorch. We train our model from scratch using an Adam optimizer with an initial learning rate of 5e-5. We set the batch size to 4 and train 1000 epochs for our multi-task model.

4.3. Evaluation metrics

In medical image analysis, landmark detection often corresponds to measuring specific anatomical or structural features of interest. Therefore, to comprehensively evaluate the performance of our landmark detection model, we utilize three evaluation metrics for different measurement tasks: Percentage of Correct Keypoints (PCK), Success Detection Rate (SDR), and Mean Absolute Error (MAE).

The PCK for landmark k stats the correctly detected rate within a specific tolerance threshold and is defined as follows:

$$PCK^k = \frac{\sum_i \delta(d_i^k \leq \tau)}{\sum_i 1}, \tag{8}$$

where δ is the indicator function, d is the radial error between the prediction and the ground truth, and τ denotes the tolerance. In this paper, considering inter-individual differences or structural changes caused by diseases, we adopt a commonly used relative value, i.e., 0.2 times the target dimension (Xu et al., 2024), as the relative evaluation threshold. Additionally, we use an absolute threshold of 2.5 mm, considering the dimensions of the measurement targets and their standard deviation (Nucifora et al., 2011; Lang et al., 2015).

The SDR records the proportion of correct measurements. Since the annotations for some measurement tasks may exhibit variability within a specific range, SDR can effectively serve as an alternative to PCK. Similar to PCK, we have selected both relative and absolute thresholds for the SDR metric (Lang et al., 2015; Yao et al., 2015).

The MAE calculates the average absolute difference between the prediction and ground truth. Therefore, we also utilize MAE to evaluate the model performance regarding target dimension measurements and keyframe detection. These evaluation metrics enable us to assess the reliability and accuracy of the proposed method in medical image analysis, ensuring its capability to capture anatomical features in both temporal and spatial contexts effectively .

4.4. Comparison with state-of-the-art

To validate the superiority of the proposed models in landmark detection, Tables 1 and 2 show the comparisons of the proposed ABHG with eight state-of-the-art models on three datasets. The comparison models include four heatmap-based models, UNet (Ronneberger et al., 2015), Hourglass network (Newell et al., 2016), HRNet (Wang et al., 2021b), and ViTPose (Xu et al., 2022); a one-hot classification model, CHH (McCouat and Voiculescu, 2022); one Bayesian-based model, U-LanD (Jafari et al., 2022); and two coarse-to-fine models DAG (Li et al., 2020) and SA-LSTM (Chen et al., 2022). In particular, we employ two encoder–decoder architectures, UNET and Hourglass, as backbones and build two multi-task models, named ABHG_U and ABHG_H, respectively, to verify the applicability of ABHG to different backbones. It is worth noting that our model in Table 2 is pruned with the keyframe branch due to the UIC dataset consisting of single-frame data.

From the results on all three datasets, we find that the ABHG built on both backbones outperforms the other eight state-of-the-art methods. The F-test and Friedman test for MAE show that the difference is statistically significant with $p < 0.05$. On the one hand, the experimental results show the advantages of our coarse-to-fine localization strategy, both for landmark localization and target dimension measurement. On the other hand, the experiments also demonstrate that the introduction of ABHG can further enhance the single-stage model, demonstrating the generalization of our model (UNet vs. ABHG_U and Hourglass vs. ABHG_H). We also note that while HRNet and ViTPose outperform on common object keypoint detection, their performance is underwhelming on sparsely annotated temporal tasks. CHH, a method that directly predicts coordinates without the aid of the Gaussian heatmap, can also yield comparable results. The qualitative results are shown in Fig. 4, where the heatmap-based methods are prone to fall into local optimization and produce larger errors. Comparatively, explicitly modeling the structural relationships among landmarks with graph models can better guarantee the relative positions. These experimental results demonstrate that our method combines the advantages of the two types of methods and adopts a coarse-to-fine architecture to obtain more accurate results.

Comparing the results on the two in-house cine datasets, we observe that the localization and measurement of aortic annulus landmarks on the PLAX dataset for each model outperforms the results on the LAA dataset for the left atrial appendage. These results reflect the differences between the two tasks, where the amplitude of the left atrial appendage is significantly larger than the aortic annulus during a cardiac cycle, and the features are relatively ambiguous, harboring greater uncertainty. Moreover, we also observed larger standard deviations on the

Table 1
Comparison of Landmark Detection with state-of-the-art methods on LAA and PLAX datasets using evaluation metrics of Percentage of Correct Keypoints (PCK) and Mean Absolute Error (MAE). MAE is expressed as mean ± SD.

Data	Model	PCK@0.2 ↑					PCK@2.5 mm ↑					MAE (mm) ↓		
		P1	P2	P3	P4	Mean	P1	P2	P3	P4	Mean	LAAO (AA)	LAAD	Mean
LAA	UNet	0.72	0.74	0.44	0.82	0.68	0.67	0.69	0.40	0.59	0.59	2.02±0.29	3.04±0.23	2.53±0.22
	Hourglass	0.59	0.66	0.42	0.79	0.62	0.50	0.60	0.37	0.41	0.47	2.22±0.17	3.39±0.22	2.81±0.16
	HRNet	0.54	0.52	0.42	0.80	0.57	0.52	0.44	0.46	0.42	0.46	1.91±0.23	3.69±0.36	2.80±0.20
	ViTPose	0.42	0.56	0.20	**0.86**	0.51	0.30	0.48	0.22	0.32	0.33	3.01±0.22	3.91±0.40	3.46 ±0.21
	CHH	**0.80**	0.77	0.38	0.82	0.69	0.65	0.68	0.42	0.61	0.59	1.96±0.28	3.03±0.46	2.50±0.37
	U-LanD	0.72	**0.80**	0.44	0.83	0.70	0.67	**0.72**	0.40	0.59	0.60	2.31±0.21	3.40±0.21	2.86±0.18
	DAG	0.54	0.69	0.31	0.74	0.57	0.41	0.60	0.25	0.30	0.39	2.32±0.24	3.48±0.68	2.90±0.43
	SA-LSTM	0.34	0.36	0.19	0.54	0.36	0.26	0.27	0.18	0.17	0.22	2.37±0.16	3.89±0.46	3.18±0.16
	ABHG_U	**0.80**	0.76	**0.49**	**0.86**	**0.73**	**0.69**	0.69	0.44	**0.63**	**0.61**	1.97±0.11	**2.95**±0.26	2.46±0.11
	ABHG_H	0.70	0.74	**0.49**	0.84	0.69	0.67	0.68	**0.45**	0.57	0.59	**1.82**±0.18	2.97±0.29	**2.40**±0.19
PLAX	UNet	0.80	0.79	–	–	0.80	0.37	0.57	–	–	0.47	1.24±0.15	–	–
	Hourglass	0.83	0.76	–	–	0.80	0.50	0.57	–	–	0.54	0.99±0.09	–	–
	HRNet	0.90	0.70	–	–	0.80	0.43	0.27	–	–	0.35	1.25±0.13	–	–
	ViTPose	0.60	0.63	–	–	0.62	0.13	0.20	–	–	0.17	2.46±0.20	–	–
	CHH	0.83	0.78	–	–	0.81	0.64	0.59	–	–	0.62	1.01±0.09	–	–
	U-LanD	0.84	0.77	–	–	0.81	**0.67**	0.57	–	–	0.62	0.88±0.14	–	–
	DAG	0.83	**0.81**	–	–	0.82	0.50	0.43	–	–	0.47	1.21±0.12	–	–
	SA-LSTM	0.73	0.70	–	–	0.72	0.23	0.30	–	–	0.27	1.33±0.12	–	–
	ABHG_U	0.83	0.80	–	–	0.82	**0.67**	**0.60**	–	–	**0.64**	**0.80**±0.13	–	–
	ABHG_H	**0.93**	0.73	–	–	**0.83**	0.60	0.57	–	–	0.62	0.83±0.12	–	–

Table 2
Comparison of Landmark Detection with state-of-the-art methods on the UIC dataset and by evaluation metrics of Success Detection Rate (SDR) and Mean Absolute Error (MAE). MAE is expressed as median (first quartile-third quartile).

Data	Model	SDR@0.2 ↑				SDR@3/1.5/1.5 mm ↑				MAE (mm) ↓		
		LVID	IVS	PW	Mean	LVID	IVS	PW	Mean	LVID	IVS	PW
UIC	UNet	0.83	0.67	0.57	0.69	0.57	0.53	0.41	0.50	2.52 (1.08-5.49)	1.44 (0.66-3.23)	1.91 (1.01-3.85)
	Hourglass	0.83	0.56	0.53	0.64	0.49	0.41	0.42	0.44	3.06 (1.44-5.97)	2.00 (0.96-3.44)	1.91 (0.96-3.45)
	HRNet	0.84	0.51	0.51	0.62	0.51	0.39	0.35	0.42	2.91 (1.33-5.58)	2.06 (1.05-3.46)	2.30 (1.10-3.91)
	ViTPose	0.73	0.37	0.35	0.48	0.35	0.24	0.25	0.28	4.75 (2.03-7.90)	3.42 (1.50-5.47)	3.55 (1.44-6.45)
	CHH	0.87	0.59	0.39	0.62	0.53	0.45	0.27	0.42	2.63 (1.15-5.69)	1.59 (0.66-3.70)	3.19 (1.30-5.15)
	U-LanD	0.87	0.65	0.59	0.70	0.56	0.48	0.42	0.49	2.48 (1.10-4.74)	1.57 (0.71-3.01)	1.83 (0.75-3.31)
	DAG	**0.90**	0.70	0.59	0.73	0.59	0.55	0.47	0.54	2.41 (0.99-4.84)	**1.38** (0.68-**2.30**)	1.65 (0.75-3.06)
	SA-LSTM	0.85	0.65	0.57	0.69	0.53	0.51	0.43	0.49	2.68 (1.23-5.24)	1.49 (0.63-2.84)	1.82 (0.88-3.51)
	ABHG_U	**0.90**	0.68	0.63	0.74	0.59	0.49	0.47	0.52	2.24 (0.88-**4.35**)	1.50 (**0.60**-2.76)	1.64 (0.79-2.93)
	ABHG_H	**0.90**	**0.75**	**0.67**	**0.77**	**0.63**	**0.59**	**0.49**	**0.57**	**2.09** (**0.87**-4.57)	1.53 (0.73-2.58)	**1.28** (**0.55-2.21**)

UIC dataset. One reason for this discrepancy is the inherent variability in the landmarks used for left ventricular measurements, which may vary along the direction parallel to the long axis of the left ventricle. Another reason is that UIC is a multicenter dataset, which introduces greater heterogeneity.

For the keyframe detection task, we compare the model uncertainty-based approach, U-LanD, and the temporal model-based approach, Dense-GRU (T. Dezaki et al., 2019). In addition to comparing the model structure, we also compare the performance difference between the proposed Order Loss and GE Loss (T. Dezaki et al., 2019), which treats the keyframe index as a global extreme. Referring to the experimental outcomes presented in Table 1, we opt for UNet as the backbone to train our multi-task model with various loss functions. We evaluate the model performance using the MAE between the detection frame and the ground truth. Notably, one of the major challenges of this work is sparse annotation, and we need to consider how to evaluate landmark detection when keyframes are detected incorrectly. Considering that detecting landmarks enables automated measurements, we also calculate the MAE of the target dimension between detected frames and labeled keyframes. This MAE can also reflect the rationality of the keyframe prediction and the accuracy of the landmark detection, i.e., the higher the MAE, the larger the landmark deviation.

Table 3 demonstrates the apparent superiority of our proposed method on both datasets, as our model not only considers the temporal features of the images but also incorporates the motion states of the anatomical landmarks. Although the U-LanD model shows comparable performance in the landmark detection task, the error in keyframe detection is large. We believe that relying solely on the uncertainty of landmarks detection to distinguish keyframes from non-keyframes is not an optimal solution, especially when the difference between frames is subtle. In addition, combining different models with loss functions demonstrates that the proposed Order Loss and multi-task architecture have great potential to recognize keyframes of interest in dynamic sequences. The T-test indicates a statistically significant difference between our multi-task model and the corresponding single-task model based on GRU, with $p < 0.05$.

Given that the ultimate goal of this study is to automate measurements in ultrasonic cine, we further utilize Bland–Altman analysis to assess the agreement between the model predictions and manual measurements. The Bland–Altman plot, displayed in Fig. 5, shows the mean (x-axis) of the two measurements and the difference (y-axis) between them. We observed that the mean differences are 0.232 mm, −0.630 mm, and −0.487 mm, respectively, and that most points fall between the Mean ±1.96 SD. The Bland–Altman analysis demonstrates the excellent accuracy and consistency of our proposed model with experienced physicians.

4.5. Ablation analysis

4.5.1. Effect of ABHG

The purpose of ABHG is twofold: to improve the accuracy of the single-stage model, enabling coarse-to-fine landmark detection, and

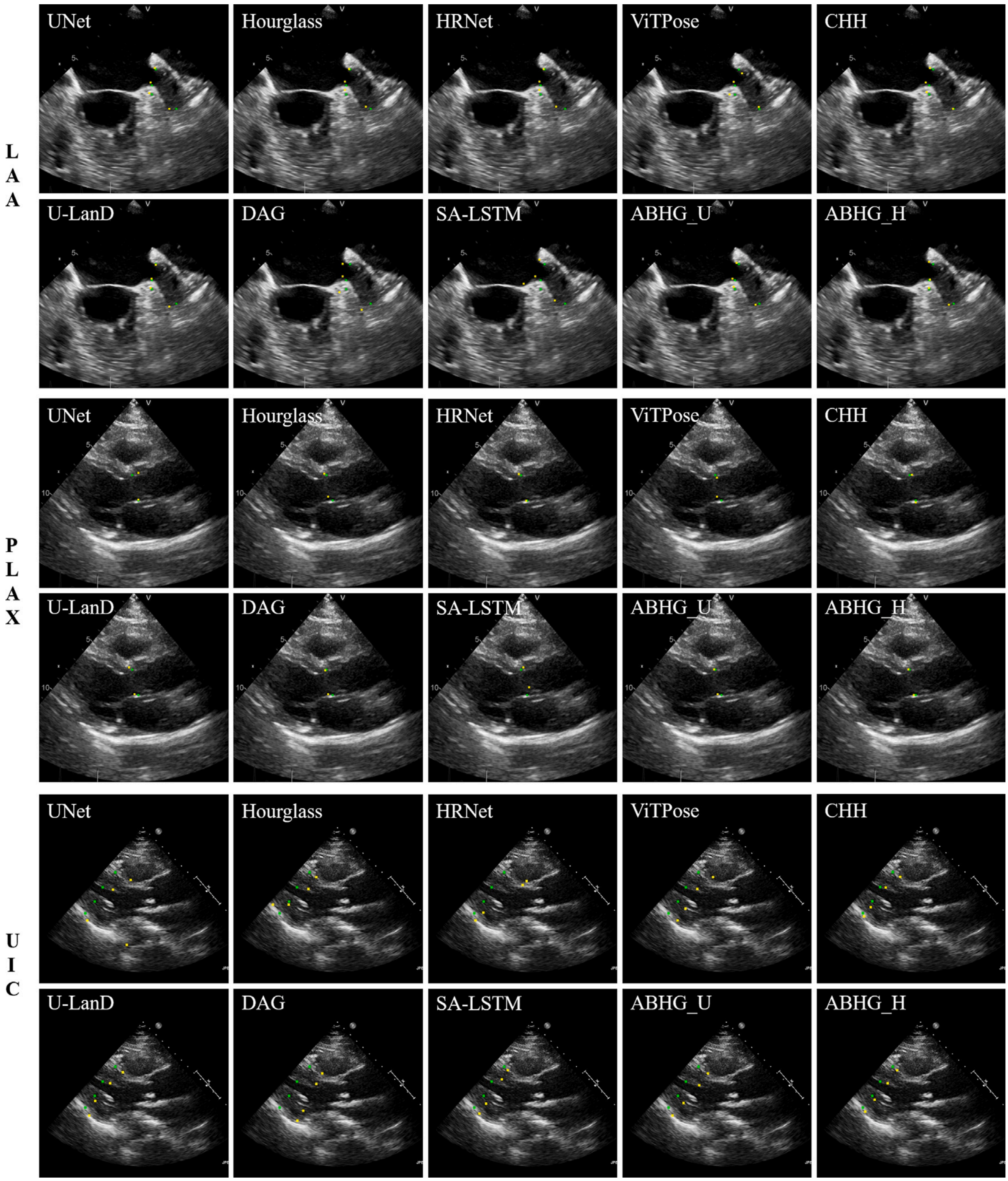

Fig. 4. Qualitative results for locating the landmarks of the left atrial appendage, aortic annulus, and left ventricle. Green points indicate ground truth, while yellow points mark the predictions.

to outperform other state-of-the-art methods. As shown in Table 1, ABHG results outperform other methods in both point accuracy and measurement error. Furthermore, ABHG achieves a 5% improvement in PCK@0.2 compared to UNet and a 7% improvement compared to Hourglass. The MAE decreases from 2.53 mm to 2.46 mm relative to UNet and from 2.81 mm to 2.40 mm relative to Hourglass, demonstrating a significant advancement in landmarks detection accuracy ($p < 0.05$). These results also highlight the generalization ability of the ABHG model based on different backbone architectures.

In addition to comparing with fine-tuning models based on graph structures like DAG and SA-LSTM, we also attempted to replace ABHG with a simple graph structure (Kipf and Welling, 2017) for comparison. As shown in Table 4, the proposed ABHG outperforms the basic GCN, particularly on the relatively challenging LAA dataset. Furthermore, it can be observed that even though the representation of the hypergraph model is more complex compared to the traditional graph model, it does not increase the number of training parameters. In fact, hypergraph can be lighter and more flexible when the node size is small.

Further, we quantitatively analyze the fine-tuning of landmarks by ABHG, as shown in Fig. 6. We introduced a random displacement to the coarse localization results, creating a new initial point for ABHG, represented by the blue point in the figure. After 20 iterations of ABHG fine-tuning, the trajectory of the red marker gradually converges to

Table 3
Comparison of Keyframe Detection with state-of-the-art methods on LAA and PLAX datasets using evaluation metrics of Mean Absolute Error (MAE). MAE is expressed as mean ± SD.

Data	Model	MAE (frame) ↓	MAE (mm) ↓	
			LAAO(AA)	LAAD
LAA	U-LanD	13.28±2.75	4.54±0.53	7.73 ±0.71
	Dense-GRU + GE Loss	6.58±0.62	–	–
	Dense-GRU + Order Loss	5.94±0.47	–	–
	Our + GE Loss	4.38±0.53	2.34±0.23	3.80±0.27
	Our + PL Loss	4.23±0.60	2.30±0.25	3.77±0.21
	Our + Order Loss	**4.08**±0.41	**2.14**±0.21	**3.72**±0.23
PLAX	U-LanD	13.73±2.10	1.56±0.44	–
	Dense-GRU + GE Loss	2.93±1.18	–	–
	Dense-GRU + Order Loss	2.37±1.40	–	–
	Our + GE Loss	1.90±1.72	1.78±0.19	–
	Our + PL Loss	1.92±1.77	1.53±0.20	–
	Our + Order Loss	**1.83**±1.52	**1.03**±0.16	–

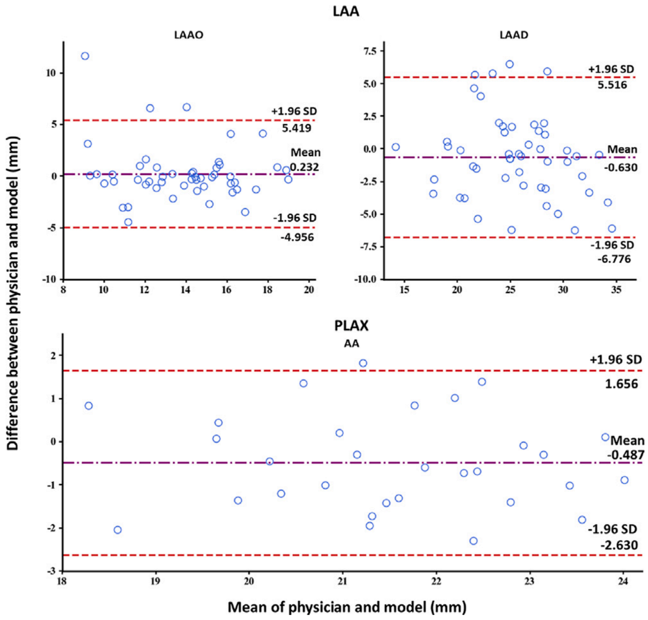

Fig. 5. Agreement between two measurements (Bland–Altman plot).

Table 4
Comparison of ABHG and GCN for Landmark Detection on LAA and PLAX Datasets. MAE is expressed as mean ±SD.

Data	Model	Parameters	PCK@0.2 ↑	PCK@2.5 mm ↑	MAE (mm) ↓
LAA	GCN	71,709	0.59	0.51	2.94±0.21
	ABHG	70,917	0.73	0.61	2.46±0.11
PLAX	GCN	71,369	0.81	0.60	1.14±0.17
	ABHG	70,521	0.82	0.64	0.80±0.13

the green ground truth, which underscores the capability of ABHG to perform localized adjustments to coordinates. However, it is essential to emphasize that the purpose of the proposed ABHG is to fine-tune the heatmap-based predictions further and that the two-stage model shares a feature encoder, thus the initial position affects the fine-tuning results.

4.5.2. Effect of multi-task learning

In this work, we propose a multi-task framework to simultaneously detect keyframes and landmarks for automatic measurements in ultrasound examinations. To verify the mutually reinforcing effect between two tasks, we trained two sub-models independently for their respective tasks based on the proposed model. In addition, we trained another model for performing two detection tasks simultaneously, which differs from the proposed model in that it does not pass the predicted coordinates to the keyframe branch. From Fig. 7, we observe that the proposed multi-task model improves substantially on both tasks, especially for the keyframe detection. However, simply combining two sub-models without establishing meaningful associations leads to sub-optimal performance.

Table 5
Accuracy of keyframe detection under varying degrees of landmark location errors. MAE is expressed as mean ±SD.

Landmark	Ground truth	Predictions	Ground truth $\pm N(10,3)$	Ground truth $\pm N(20,3)$	All zeros
MAE (frame)	4.16±0.35	4.08±0.41	4.78±0.43	4.64±0.50	5.34±0.51

Furthermore, we evaluate the effect of landmark accuracy on keyframe identification by introducing varying degrees of noise to the landmark coordinates of the keyframe passed to the keyframe branch. The results shown in Table 5 indicate a decrease in keyframe accuracy when Gaussian noise with means of 10 and 20 pixels is applied to the landmark coordinates. Moreover, the accuracy notably declines when the passed landmark coordinates are all zeros. Interestingly, even with ground truth provided, the accuracy does not increase. This behavior is attributed to our multi-task model, which, trained on sparse data, predicts keyframes based on landmarks and features extracted from image sequences by the shared encoder–decoder. In summary, we argue that the landmark coordinates provide essential information for identifying the keyframe for these sparsely annotated training tasks. Simultaneously, the supervision of frame likelihoods makes landmark detection more concerned with subtle changes in the keyframe.

4.5.3. Effect of MC Dropout

In this section, we thoroughly investigate the impact of MC Dropout (Gal and Ghahramani, 2016) on the performance and reliability of our model. We conducted a study with five different dropout rates to evaluate the performance of left atrial appendage dimension measurements using the MAE and PCK. As illustrated in Fig. 8(a), we observe a remarkable consistency in the median of MAE after applying ABHG fine-tuning, regardless of the inclusion of MC Dropout. The median of MAE consistently hovered within the range of 1.5 mm to 2 mm across all tested dropout rates, and no statistically significant differences are observed among the models ($p > 0.05$).

Further, We count the percentage of test samples with different correct detection points at different drop rates. As illustrated in Fig. 8(b), our analysis revealed the following: (1) without the integration of MC Dropout, 2% of the test data exhibited complete detection failure, i.e., the errors for all four landmarks surpassed the tolerance. In contrast, (2) with the introduction of MC Dropout, the model demonstrated a consistent trend of improved detection accuracy, notably leading to a noticeable increase in the proportion of samples with all four landmarks correctly detected. These results illustrate the potential impact of MC Dropout on improving detection accuracy under sub-optimal image quality while also underscoring its effectiveness as a viable approach for landmark detection in dynamic data.

Similarly, we also incorporate MC Dropout in each gate of GRU and achieve an improvement of 0.8 frames in detection accuracy. Thus, we firmly contend that MC Dropout effectively mitigates the inherent uncertainty within keyframe and landmark detection tasks for ultrasonic cine.

4.5.4. Hyper-parameters selection

In this section, we investigate the impact of two important hyper-parameters in ABHG on the accuracy of landmark detection. Specifically, we crop a specific size patch from the shallow features of the original image and the encoder to construct the local features of the

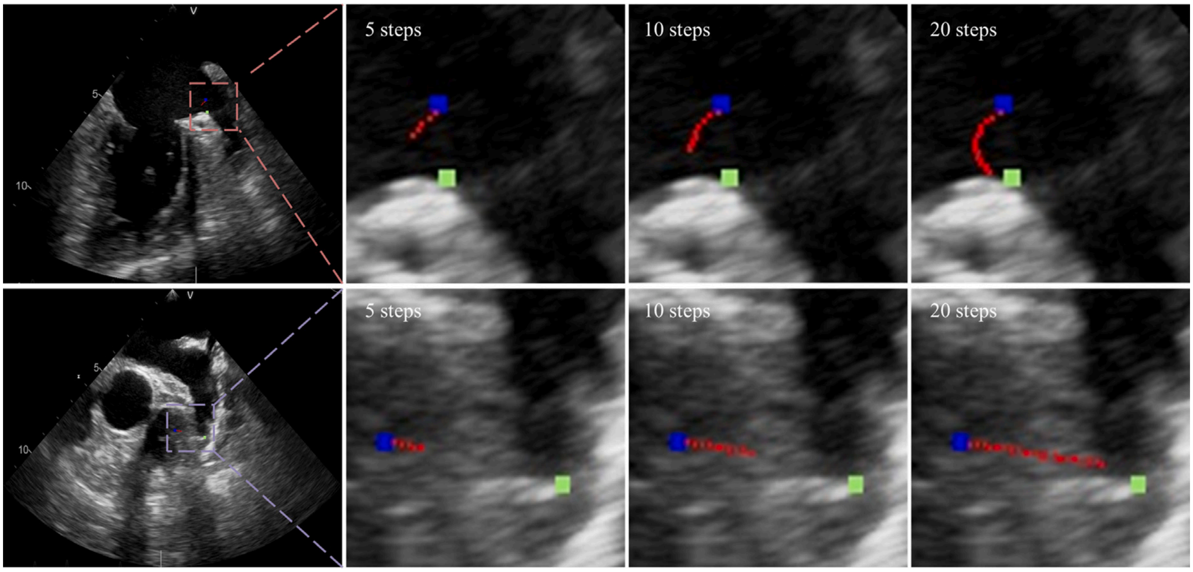

Fig. 6. Landmark trajectories over 20 iterations of ABHG. A random initial position is marked in blue, the ground truth in green, and the updating trajectory in red.

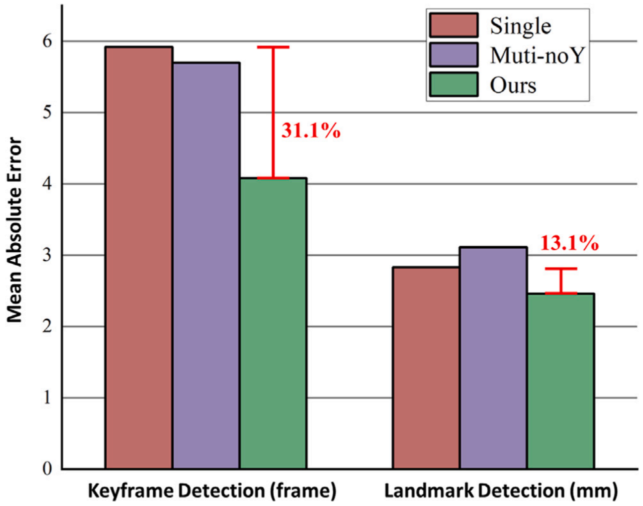

Fig. 7. Effectiveness assessment of multi-task learning. The bar chart shows the comparison of the mean absolute errors among the single-task, simple multi-task, and proposed multi-task models on the two detection tasks.

ABHG nodes, and the patch size not only affects the receptive field of the hypergraph nodes but also determines the upper bound of the offset for fine-tuning the coordinates. Table 6 shows that selecting a smaller patch size is more effective for precisely fine-tuning the landmark coordinates, therefore we choose the 8 × 8 patches to construct the local features for the nodes of ABHG.

In addition, the number of hyper-edges determines the expressiveness and performance of the hypergraph. As the number of hyper-edges increases, the model complexity and computations also increase, and the model is more prone to overfitting. Table 6 shows that when the number of hyper-edges increases to 32, the mean PCK obviously starts to decrease. We adopt 16 hyperedges to construct ABHG because it achieves the highest results under a strict evaluation metric, PCK@2.5 mm. Overall, the patch size and the number of hyperedges are critical factors in constructing hypergraphs, which need to be determined in combination with specific application and data characteristics.

5. Conclusion and discussion

In this article, we propose a Bayesian multi-task network to tackle the challenges of keyframe and landmark detection in ultrasonic cine with high uncertainty. Following the coarse-to-fine detection architecture, we propose ABHG to explicitly model the structural relationship of landmarks and fine-tune the prediction of heatmap-based regression. To overcome the potential limitations arising from information propagation and overfitting due to the scale of the hypergraph, we implement adaptive expansion within the 8-neighborhood directions of hyper-nodes. Furthermore, we establish a synergistic relationship between the receptive field of hyper-nodes and the landmark offset and introduce MC Dropout during testing to further improve the landmark localization accuracy.

To achieve a more versatile keyframe detection model, we devise the Order Loss, which captures the relative magnitude of likelihoods across frames in a cine sequence. Our experimental results on transthoracic and multi-view transesophageal echocardiography datasets demonstrated the effectiveness of the proposed approach. By jointly addressing keyframe and landmark detection, our method outperformed state-of-the-art models on both tasks, showcasing the significance of leveraging their inherent correlations.

Nevertheless, challenges remain in refining our model. At this stage, the keyframe branch in our model adopts an offline model, Bi-GRU, which performs predictions after capturing the complete sequence. While this design enables comprehensive modeling of global information, it also incurs high computational costs and memory consumption. As a follow-up, developing a real-time model to assess keyframe likelihood is a promising direction. Furthermore, although we introduce MC Dropout in both branches of our multi-task model and achieve accuracy gains, we have not yet identified a quantitative method that accurately measures model uncertainty for dynamic sequences. Future work could involve investigating advanced uncertainty quantification techniques to provide more nuanced insights into the temporal model reliability.

In conclusion, our proposed Bayesian multi-task network demonstrates the potential of jointly addressing keyframe and landmark detection in ultrasound cine. By embracing the underlying correlations of two tasks and introducing innovative architectural components, we improve the accuracy of both tasks, offering promising applications in clinical settings and laying the groundwork for further research.

Table 6
Hyper-parameters impact assessment on landmark detection model on LAA dataset using evaluation metrics of Percentage of Correct Keypoints (PCK) and Mean Absolute Error (MAE). MAE is expressed as mean ±SD.

Hyper-parameters	Configs	PCK@0.2 ↑					PCK@2.5 mm ↑					MAE (mm) ↓		
		P1	P2	P3	P4	Mean	P1	P2	P3	P4	Mean	LAAO	LAAD	Mean
Patch Size	4	**0.80**	0.78	0.46	0.88	0.73	**0.80**	**0.78**	0.48	0.62	**0.67**	2.38±0.13	**2.96**±0.31	2.67±0.23
	8	**0.80**	0.74	**0.52**	**0.90**	**0.74**	**0.80**	0.64	0.48	**0.66**	0.65	**2.21**±0.14	3.10±0.28	**2.66**±0.19
	16	0.78	**0.82**	0.50	0.82	0.73	**0.80**	0.76	**0.52**	0.56	0.66	2.65±0.19	3.32±0.24	2.99±0.22
	32	0.78	0.78	0.40	**0.90**	0.72	0.78	0.70	0.42	0.62	0.63	2.37±0.21	3.98±0.33	3.18±0.27
Number of Hyper-edge	4	0.76	**0.80**	0.48	**0.90**	**0.74**	**0.80**	0.72	0.38	**0.66**	0.64	2.85±0.28	**2.92**±0.33	2.89±0.31
	8	**0.80**	0.74	**0.52**	**0.90**	**0.74**	**0.80**	0.64	0.48	**0.66**	0.65	2.21±0.18	3.10±0.21	2.66±0.20
	16	**0.80**	**0.80**	0.48	0.84	0.73	**0.80**	**0.82**	**0.50**	0.60	**0.68**	**1.95**±0.14	3.25±0.23	2.60±0.17
	32	0.78	**0.80**	0.42	0.88	0.72	0.76	0.76	0.38	0.64	0.64	1.96±0.19	3.09±0.31	**2.53**±0.27
	64	**0.80**	0.66	0.38	0.88	0.68	0.76	0.60	0.42	**0.66**	0.61	2.18±0.22	3.16±0.27	2.67±0.25

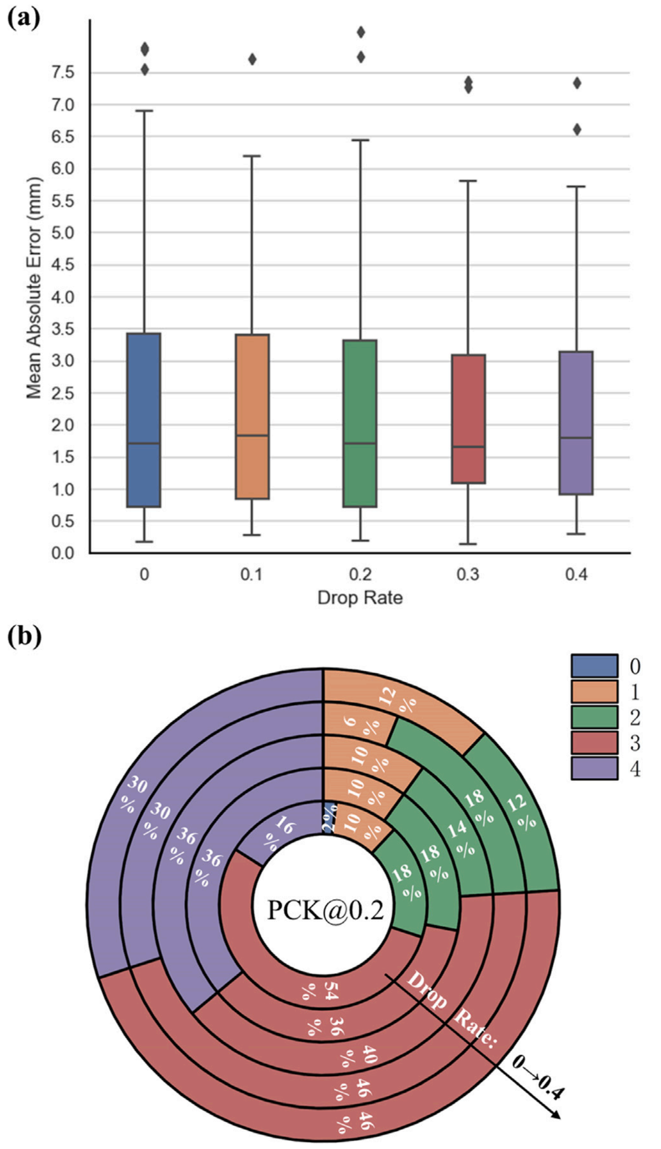

Fig. 8. Effectiveness assessment of Monte Carlo Dropout. (a) Illustration of the MAE for measurements of left atrial appendage at different drop rates. (b) The doughnut chart shows the proportion of samples correctly detecting varying numbers of landmarks at different drop rates using the PCK@0.2 metric. Different colors indicate different proportions of correctly detected landmarks. 0 indicates that all landmark coordinate errors are not within tolerance, while 4 indicates that all four landmark coordinate errors are within tolerance.

CRediT authorship contribution statement

Yong Feng: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jinzhu Yang:** Writing – review & editing, Validation, Supervision, Resources, Funding acquisition, Conceptualization. **Meng Li:** Validation, Resources, Investigation, Data curation. **Lingzhi Tang:** Project administration, Methodology, Data curation. **Song Sun:** Visualization, Investigation, Data curation. **Yonghuai Wang:** Writing – review & editing, Validation, Supervision, Resources, Data curation.

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.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Nos. U21A20387, U22A2022), the National “111” Project (No. B16009), and the Liaoning Province “Applied Basic Research Plan” project (No. 2022020421-JH2/1013).

References

Alansary, A., et al., 2019. Evaluating reinforcement learning agents for anatomical landmark detection. Med. Image Anal. 53, 156–164. http://dx.doi.org/10.1016/j.media.2019.02.007.

Chen, R., et al., 2022. Structure-aware long short-term memory network for 3D cephalometric landmark detection. IEEE Trans. Med. Imaging 41 (7), 1791–1801. http://dx.doi.org/10.1109/TMI.2022.3149281.

Cho, K., et al., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proc. EMNLP. pp. 1724–1734. http://dx.doi.org/10.3115/v1/d14-1179.

Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, URL: http://arxiv.org/abs/1412.3555.

Ciusdel, C., et al., 2020. Deep neural networks for ECG-free cardiac phase and end-diastolic frame detection on coronary angiographies. Comput. Med. Imag. Grap. 84, 101749. http://dx.doi.org/10.1016/j.compmedimag.2020.101749.

Dai, W., Li, X., Ding, X., Cheng, K.-T., 2023. Cyclical self-supervision for semi-supervised ejection fraction prediction from echocardiogram videos. IEEE Trans. Med. Imaging 42 (5), 1446–1461. http://dx.doi.org/10.1109/TMI.2022.3229136.

Dezaki, F.T., et al., 2017. Deep residual recurrent neural networks for characterisation of cardiac cycle phase from echocardiograms. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 100–108. http://dx.doi.org/10.1007/978-3-319-67558-9_12.

Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., Cardoso, M.J., 2018. Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.. pp. 691–699. http://dx.doi.org/10.1007/978-3-030-00928-1_78.

Elman, J.L., 1990. Finding structure in time. Cogn. Sci. 14 (2), 179–211. http://dx.doi.org/10.1016/0364-0213(90)90002-E.

Feng, Y., You, H., Zhang, Z., Ji, R., Gao, Y., 2019. Hypergraph neural networks. In: Proc. AAAI. pp. 3558–3565. http://dx.doi.org/10.1609/aaai.v33i01.33013558.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proc. ICML, vol. 48, pp. 1050–1059, URL: https://proceedings.mlr.press/v48/gal16.html.

Geyer, C.J., 1992. Practical markov chain monte carlo. Stat. Sci. 473–483, URL: http://www.jstor.org/stable/2246094.

Gilbert, A., Holden, M., Eikvil, L., Aase, S.A., Samset, E., McLeod, K., 2019. Automated left ventricle dimension measurement in 2D cardiac ultrasound via an anatomically meaningful CNN approach. In: Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis. pp. 29–37. http://dx.doi.org/10.1007/978-3-030-32875-7_4.

Graves, A., 2011. Practical variational inference for neural networks. In: Proc. Adv. Neural Inf. Process. Syst., vol. 24, URL: https://proceedings.neurips.cc/paper_files/paper/2011/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf.

He, Z., Li, W., Zhang, T., Yuan, Y., 2023. H2GM: A hierarchical hypergraph matching framework for brain landmark alignment. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. pp. 548–558. http://dx.doi.org/10.1007/978-3-031-43999-5_52.

Hiasa, Y., Otake, Y., Takao, M., Ogawa, T., Sugano, N., Sato, Y., 2020. Automated muscle segmentation from clinical CT using Bayesian U-net for personalized musculoskeletal modeling. IEEE Trans. Med. Imaging 39 (4), 1030–1040. http://dx.doi.org/10.1109/TMI.2019.2940555.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. http://dx.doi.org/10.1162/NECO.1997.9.8.1735.

Howard, J.P., Stowell, C.C., Cole, G.D., Ananthan, K., Demetrescu, C.D., Pearce, K., Rajani, R., Sehmi, J., Vimalesvaran, K., Kanaganayagam, G.S., et al., 2021. Automated left ventricular dimension assessment using artificial intelligence developed and validated by a UK-wide collaborative. Circ.: Cardiovasc. Imaging 14 (5), e011951. http://dx.doi.org/10.1161/circimaging.120.011951.

Huang, R., et al., 2022. Extracting keyframes of breast ultrasound video using deep reinforcement learning. Med. Image Anal. 80, http://dx.doi.org/10.1016/j.media.2022.102490, Art. no. 102490.

Jafari, M.H., Woudenberg, N.V., Luong, C., Abolmaesumi, P., Tsang, T., 2021. Deep Bayesian image segmentation for a more robust ejection fraction estimation. In: Proc ISBI. pp. 1264–1268. http://dx.doi.org/10.1109/ISBI48211.2021.9433781.

Jafari, M.H., et al., 2022. U-LanD: Uncertainty-driven video landmark detection. IEEE Trans. Med. Imaging 41 (4), 793–804. http://dx.doi.org/10.1109/TMI.2021.3123547.

Jahren, T.S., Steen, E.N., Aase, S.A., Solberg, A.H.S., 2020. Estimation of end-diastole in cardiac spectral Doppler using deep learning. IEEE Trans. Ultrason. Ferroelectr. Freq. Control 67 (12), 2605–2614. http://dx.doi.org/10.1109/TUFFC.2020.2995118.

Jin, H., Che, H., Chen, H., 2023. Unsupervised domain adaptation for anatomical landmark detection. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. pp. 695–705. http://dx.doi.org/10.1007/978-3-031-43907-0_66.

Jospin, L.V., Laga, H., Boussaid, F., Buntine, W., Bennamoun, M., 2022. Hands-on Bayesian neural networks—A tutorial for deep learning users. IEEE Comput. Intell. Mag. 17 (2), 29–48. http://dx.doi.org/10.1109/MCI.2022.3155327.

Kendall, A., Badrinarayanan, V., Cipolla, R., 2017. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In: Proc. BMVC. http://dx.doi.org/10.5244/C.31.57.

Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. In: Proc. ICLR. URL: https://openreview.net/forum?id=SJU4ayYgl.

Kong, B., Zhan, Y., Shin, M., Denny, T., Zhang, S., 2016. Recognizing end-diastole and end-systole frames via deep temporal regression network. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.. pp. 264–272. http://dx.doi.org/10.1007/978-3-319-46726-9_31.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Proc. Adv. Neural Inf. Process. Syst., vol. 25, URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proc. Adv. Neural Inf. Process. Syst., vol. 30, URL: https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Lane, E.S., et al., 2021. Multibeat echocardiographic phase detection using deep neural networks. Comput. Biol. Med. 133, http://dx.doi.org/10.1016/j.compbiomed.2021.104373, Art. no. 104373.

Lang, R.M., et al., 2015. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American society of echocardiography and the European association of cardiovascular imaging. J. Am. Soc. Echocardiogr. 28 (1), 1–39. http://dx.doi.org/10.1016/j.echo.2014.10.003.

Lang, Y., et al., 2020. Automatic localization of landmarks in craniomaxillofacial CBCT images using a local attention-based graph convolution network. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.. pp. 817–826. http://dx.doi.org/10.1007/978-3-030-59719-1_79.

Leitner, C., et al., 2022. A human-centered machine-learning approach for muscle-tendon junction tracking in ultrasound images. IEEE Trans. Biomed. Eng. 69 (6), 1920–1930. http://dx.doi.org/10.1109/TBME.2021.3130548.

Li, W., et al., 2020. Structured landmark detection via topology-adapting deep graph learning. In: Proc. ECCV. pp. 266–283. http://dx.doi.org/10.1007/978-3-030-58545-7_16.

Lu, G., Zhang, Y., Kong, Y., Zhang, C., Coatrieux, J.-L., Shu, H., 2022. Landmark localization for cephalometric analysis using multiscale image patch-based graph convolutional networks. IEEE J. Biomed. Health Inf. 26 (7), 3015–3024. http://dx.doi.org/10.1109/JBHI.2022.3157722.

Maraci, M., Bridge, C., Napolitano, R., Papageorghiou, A., Noble, J., 2017. A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat. Med. Image Anal. 37, 22–36. http://dx.doi.org/10.1016/j.media.2017.01.003.

McCouat, J., Voiculescu, I., 2022. Contour-hugging heatmaps for landmark detection. In: Proc. CVPR. pp. 20565–20573. http://dx.doi.org/10.1109/CVPR52688.2022.01994.

Mitchell, C., et al., 2019. Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the American society of echocardiography. J. Am. Soc. Echocardiogr. 32 (1), 1–64. http://dx.doi.org/10.1016/j.echo.2018.06.004.

Mokhtari, M., Mahdavi, M., Vaseli, H., Luong, C., Abolmaesumi, P., Tsang, T.S.M., Liao, R., 2023. EchoGLAD: Hierarchical graph neural networks for left ventricle landmark detection on echocardiograms. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. pp. 227–237. http://dx.doi.org/10.1007/978-3-031-43901-8_22.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation. In: Proc. ECCV. pp. 483–499. http://dx.doi.org/10.1007/978-3-319-46484-8_29.

Noothout, J.M.H., et al., 2020. Deep learning-based regression and classification for automatic landmark localization in medical images. IEEE Trans. Med. Imaging 39 (12), 4011–4022. http://dx.doi.org/10.1109/TMI.2020.3009002.

Nucifora, G., Faletra, F.F., Regoli, F., Pasotti, E., Pedrazzini, G., Moccetti, T., Auricchio, A., 2011. Evaluation of the left atrial appendage with real-time 3-dimensional transesophageal echocardiography: implications for catheter-based left atrial appendage closure. Circ.: Cardiovasc. Imaging 4 (5), 514–523. http://dx.doi.org/10.1161/circimaging.111.963892.

Oh, K., Oh, I.-S., Le, V.N.T., Lee, D.-W., 2021. Deep anatomical context feature learning for cephalometric landmark detection. IEEE J. Biomed. Health Inf. 25 (3), 806–817. http://dx.doi.org/10.1109/JBHI.2020.3002582.

Ouyang, D., et al., 2020. Video-based AI for beat-to-beat assessment of cardiac function. Nature 580 (7802), 252–256. http://dx.doi.org/10.1038/s41586-020-2145-8.

Pu, B., Zhu, N., Li, K., Li, S., 2021. Fetal cardiac cycle detection in multi-resource echocardiograms using hybrid classification framework. Future Gener. Comput. Syst. 115, 825–836. http://dx.doi.org/10.1016/j.future.2020.09.014.

Quan, Q., Yao, Q., Li, J., Zhou, S.K., 2022. Which images to label for few-shot medical landmark detection? In: Proc. CVPR. pp. 20574–20584. http://dx.doi.org/10.1109/CVPR52688.2022.01995.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.. pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Rueda, S., Alcañiz, M., 2006. An approach for the automatic cephalometric landmark detection using mathematical morphology and active appearance models. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. pp. 159–166. http://dx.doi.org/10.1007/11866565_20.

Schobs, L.A., Swift, A.J., Lu, H., 2023. Uncertainty estimation for heatmap-based landmark localization. IEEE Trans. Med. Imaging 42 (4), 1021–1034. http://dx.doi.org/10.1109/TMI.2022.3222730.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: A unified embedding for face recognition and clustering. In: Proc. CVPR. pp. 815–823. http://dx.doi.org/10.1109/CVPR.2015.7298682.

Sedai, S., Roy, P.K., Garnavi, R., 2015. Right ventricle landmark detection using multiscale HOG and random forest classifier. In: Proc. IEEE 12th Int. Symp. Biomed. Imag.. ISBI, pp. 814–818. http://dx.doi.org/10.1109/ISBI.2015.7163996.

Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y., 2018. Integral human pose regression. In: Proc. ECCV. pp. 536–553. http://dx.doi.org/10.1007/978-3-030-01231-1_33.

T. Dezaki, F., et al., 2019. Cardiac phase detection in echocardiograms with densely gated recurrent neural networks and global extrema loss. IEEE Trans. Med. Imaging 38 (8), 1821–1832. http://dx.doi.org/10.1109/TMI.2018.2888807.

Tang, L., et al., 2024. A new automated prognostic prediction method based on multi-sequence magnetic resonance imaging for hepatic resection of colorectal cancer liver metastases. IEEE J. Biomed. Health Inf. 28 (3), 1528–1539. http://dx.doi.org/10.1109/JBHI.2024.3350247.

Toshev, A., Szegedy, C., 2014. DeepPose: Human pose estimation via deep neural networks. In: Proc. CVPR. pp. 1653–1660. http://dx.doi.org/10.1109/CVPR.2014.214.

Tripathi, A., et al., 2023. Unsupervised landmark detection and classification of lung infection using transporter neural networks. Comput. Biol. Med. 152, 106345. http://dx.doi.org/10.1016/j.compbiomed.2022.106345.

Wang, Z., Shi, J., Hao, X., Wen, K., Jin, X., An, H., 2021a. Simultaneous right ventricle end-diastolic and end-systolic frame identification and landmark detection on echocardiography. In: Proc. EMBC IEEE. pp. 3916–3919. http://dx.doi.org/10.1109/EMBC46164.2021.9630310.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B., 2021b. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43 (10), 3349–3364. http://dx.doi.org/10.1109/TPAMI.2020.2983686.

Wang, Y., et al., 2022. Key-frame guided network for thyroid nodule recognition using ultrasound videos. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.. pp. 238–247. http://dx.doi.org/10.1007/978-3-031-16440-8_23.

Xu, J., Yang, B., Chen, Z., Lv, P., Zhang, S., Luo, J., 2024. Ske-Fi: Estimating hand poses via RF vision under low contrast and occlusion. IEEE Internet Things J. 11 (4), 6412–6425. http://dx.doi.org/10.1109/JIOT.2023.3312316.

Xu, Y., Zhang, J., ZHANG, Q., Tao, D., 2022. ViTPose: Simple vision transformer baselines for human pose estimation. In: Proc. Adv. Neural Inf. Process. Syst., vol. 35, pp. 38571–38584, URL: http://papers.nips.cc/paper_files/paper/2022/hash/fbb10d319d44f8c3b4720873e4177c65-Abstract-Conference.html.

Xu, Z., et al., 2018. Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images. In: Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.. pp. 711–719. http://dx.doi.org/10.1007/978-3-030-00934-2_79.

Xu, J., et al., 2021. Hip landmark detection with dependency mining in ultrasound image. IEEE Trans. Med. Imaging 40 (12), 3762–3774. http://dx.doi.org/10.1109/TMI.2021.3097355.

Yao, G.-H., Deng, Y., Liu, Y., Xu, M.-J., Zhang, C., Deng, Y.-B., Ren, W.-D., Li, Z.-A., Tang, H., Zhang, Q.-B., et al., 2015. Echocardiographic measurements in normal Chinese adults focusing on cardiac chambers and great arteries: a prospective, nationwide, and multicenter study. J. Am. Soc. Echocardiogr. 28 (5), 570–579. http://dx.doi.org/10.1016/j.echo.2015.01.022.

Zhang, C., Bütepage, J., Kjellström, H., Mandt, S., 2019. Advances in variational inference. IEEE Trans. Pattern Anal. Mach. Intell. 41 (8), 2008–2026. http://dx.doi.org/10.1109/TPAMI.2018.2889774.

Zhang, R., Qin, B., Zhao, J., Zhu, Y., Lv, Y., Ding, S., 2023. Locating X-ray coronary angiogram keyframes via long short-term spatiotemporal attention with image-to-patch contrastive learning. IEEE Trans. Med. Imaging 1. http://dx.doi.org/10.1109/TMI.2023.3286859.

Zhou, D., Huang, J., Schölkopf, B., 2006. Learning with hypergraphs: Clustering, classification, and embedding. In: Proc. Adv. Neural Inf. Process. Syst., vol. 19, URL: https://proceedings.neurips.cc/paper_files/paper/2006/file/dff8e9c2ac33381546d96deea9922999-Paper.pdf.

Zou, X., Zhong, S., Yan, L., Zhao, X., Zhou, J., Wu, Y., 2019. Learning robust facial landmark detection via hierarchical structured ensemble. In: Proc. ICCV. pp. 141–150. http://dx.doi.org/10.1109/ICCV.2019.00023.