·技术方法·

一种超声心动图关键帧智能检测方法

杜 悦[1],史中青[2],戚占如[2],曾子炀[3],郭冠军[2],姚 静[2],罗守华[3],顾 宁[1,4*]

[1]南京医科大学生物医学工程与信息学院,江苏 南京 211166;[2]南京大学医学院附属鼓楼医院超声医学科,江苏 南京 210008;[3]东南大学生物科学与医学工程学院,江苏 南京 210096;[4]南京大学医学院,江苏 南京 210093

[摘 要] **目的**:探讨基于深度学习(deep learning,DL)的ResNet+VST模型在超声心动图关键帧智能检测方面的可行性。**方法**:选取南京大学医学院附属鼓楼医院超声医学科采集的663个动态图像含心尖二腔(apical two chambers,A2C)、心尖三腔(apical three chambers,A3C)与心尖四腔(apical four chambers,A4C)3类临床检查常用切面以及EchoNet-Dynamic公开数据集中280个A4C切面动态图像,分别建立南京鼓楼医院数据集与EchoNet-Dynamic-Tiny数据集,各类别图像按4∶1方式划分为训练集和测试集,进行ResNet+VST模型的训练以及与多种关键帧检测模型的性能对比,验证ResNet+VST模型的先进性。**结果**:ResNet+VST模型能够更准确地检测心脏舒张末期(end-diastole,ED)与收缩末期(end-systole,ES)图像帧。在南京鼓楼医院数据集上,模型对A2C、A3C和A4C切面数据的ED预测帧差分别为1.52±1.09、1.62±1.43、1.27±1.17,ES预测帧差分别为1.56±1.16、1.62±1.43、1.45±1.38;在EchoNet-Dynamic-Tiny数据集上,模型对A4C切面数据的ED预测帧差为1.62±1.26,ES预测帧差为1.71±1.18,优于现有相关研究。此外,ResNet+VST模型有良好的实时性表现,在南京鼓楼医院数据集与EchoNet-Dynamic-Tiny数据集上,基于GTX 3090Ti GPU对16帧的超声序列片段推理的平均耗时分别为21 ms与10 ms,优于以长短期记忆单元(long short-term memory,LSTM)进行时序建模的相关研究,基本满足临床即时处理的需求。**结论**:本研究提出的ResNet+VST模型在超声心动图关键帧检测的准确性、实时性方面,相较于现有研究有更出色的表现,该模型原则上可推广到任何超声切面,有辅助超声医师提升诊断效率的潜力。

[关键词] 超声心动图;关键帧;深度学习

[中图分类号] R445.1 [文献标志码] A [文章编号] 1007-4368(2024)02-253-10

doi:10.7655/NYDXBNSN230743

An intelligent detection method of key frame in echocardiography

DU Yue[1],SHI Zhongqing[2],QI Zhanru[2],ZENG Ziyang[3],GUO Guanjun[2],YAO Jing[2],LUO Shouhua[3],GU Ning[1,4*]

[1]School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166; [2]Department of Ultrasound, Affiliated Drum Tower Hospital, Medical School, Nanjing University, Nanjing 210008; [3]School of Biological Sciences and Medical Engineering, Southeast University, Nanjing 210096; [4]Medical School of Nanjing University, Nanjing 210093, China

[**Abstract**] **Objective**: To explore the feasibility of using ResNet+VST model based on deep learning(DL)for intelligent detection of key frames in echocardiography. **Methods**: A total of 663 dynamic images including apical two chambers(A2C), apical three chambers(A3C), and apical four chambers(A4C), which are commonly used clinical examination views, were collected from the Department of Ultrasound Medicine at Drum Tower Hospital, Nanjing University Medical School. Additionally, 280 dynamic A4C images from the EchoNet-Dynamic public dataset were selected. Two datasets were established: the Nanjing Drum Tower Hospital dataset and the EchoNet-Dynamic-Tiny dataset. The images in each category were divided into training set and testing sets in a 4∶1 ratio. The ResNet+VST model was trained and its performance was compared with other key frame detection models to verify the its superiority. **Results**: The ResNet+VST model can detect the end-diastolic(ED)and end-systolic(ES)image frames of the heart more accurately. On the Nanjing Drum Tower Hospital dataset, the model achieved ED frame prediction differences of 1.52±1.09, 1.62±1.43, and 1.27±1.17 for A2C, A3C, and A4C views, respectively, and ES frame prediction differences of 1.56±1.16, 1.62±1.43, and 1.45±1.38, respectively. On

[基金项目] 江苏省重点研发计划(BE2022828);江苏省前沿引领技术基础研究专项(BK20222002)

*通信作者(Corresponding author),E-mail:guning@nju.edu.cn

the EchoNet-Dynamic-Tiny dataset, the model achieved an ED frame prediction difference of 1.62±1.26 and an ES frame prediction difference of 1.71±1.18, outperforming existing related studies. Furthermore, the ResNet+VST model exhibited good real-time performance, with average inference times of 21 ms and 10 ms for 16-frame ultrasound sequences on the Nanjing Drum Tower Hospital dataset and the EchoNet-Dynamic-Tiny dataset, respectively, using the GTX 3090Ti GPU. This performance was superior to related studies that used long short-term memory (LSTM) for temporal modeling and met the requirements for clinical real-time processing. **Conclusion**: The proposed ResNet+VST model demonstrates superior accuracy and real-time performance in the detection of key frames in echocardiography compared to existing research. In principle, this model can be applied to any ultrasound view and has the potential to assist ultrasound physicians in improving diagnostic efficiency.

[**Key words**] echocardiography; key frame; deep learning

[J Nanjing Med Univ, 2024, 44(02): 253-262]

超声心动图是一种无创、安全的医学成像方式，被广泛应用于心脏病患者的诊断和治疗[1]。超声心动图关键帧检测，通常指超声心动图动态图像舒张末期(end-diastole, ED)和收缩末期(end-systole, ES)帧的检测，是超声心动图检查的必要步骤，也是心脏大小量化、功能评价的重要基础。超声心动图中，ED帧可定义为二尖瓣(mitral valve, MV)关闭后第1帧、心动周期中左心室(left ventricle, LV)径线或容量最大的停帧，ES帧可定义为主动脉瓣(aortic valve, AV)关闭后的第1帧、心动周期中LV径线或容量最小的停帧，分别对应心电图(electrocardiogram, ECG)中R波波峰与T波终点的相应帧[2-3](图1)。目前临床场景下关键帧的识别，主要依靠超声医师肉眼观测超声心动图中LV容积或ECG波形，存在人工成本高、操作者经验依赖性高、可重复性差的问题。因此，实现对超声心动图中ED和ES帧的高精度自动检测具有重要的意义。

研究人员在超声心动图ED、ES帧自动检测方面做出了许多努力，早期研究中最常见的方法是LV分割法，即基于心脏超声图像中LV的分割结果计算LV面积确定ED和ES帧[4-7]。然而，心脏超声图像的LV分割需要复杂的预处理步骤，且超声图像具有信噪比低、边缘模糊等特性，容易导致分割效果欠佳，进而影响ED和ES帧的检测结果。近年来，深度学习(deep learning, DL)在各类自然图像处理任务中表现出优异性能，因此被广泛应用于超声心动图的图像处理与分析中[8-12]。一些研究将DL引入超声心动图关键帧检测领域，结合卷积神经网络(convolutional neural network, CNN)和循环神经网络(recurrent neural network, RNN)检测ED、ES帧。Dezaki等[13]借鉴了Kong等[14]在MRI中ED、ES帧检测的研究，结合ResNet[15]与长短期记忆单元(long short-term memory, LSTM)[16]提出了深度残差递归神经网络，以提取固定长度心脏超声图像序列的时空特征，实现ED和ES帧的检测，但需要对输入的超声序列进行预处理，以分离单个心动周期。Taheri等[17]对此进行了改进，将可变长度的超声序列输入到结合了DenseNet和门控单元的模块中，并提出了全局极值损失函数进一步提高ED、ES帧检测性能，然而输入的视频仍只能包含1个心动周期，导致检测结果存在偏差。Fiorito等[18]将3D CNN与LSTM的混合模型应用于超声心动图视频的时空特征提取，对各帧进行舒张期和收缩期分类，将ED和ES帧确定为两种状态之间的切换帧，可用于任意长度的序列，但仍只能检测包含1对ED和ES帧的视频，对于视频中其他ED帧，需要依赖QRS复杂波进行预测。Lane等[19]结合ResNet与LSTM提取超声序列的时空信息，证明了DL技术用于包含多个心动周期的任意长度超声序列ED和ES帧识别的可行性，但其计算上相对复杂，推理耗时长。

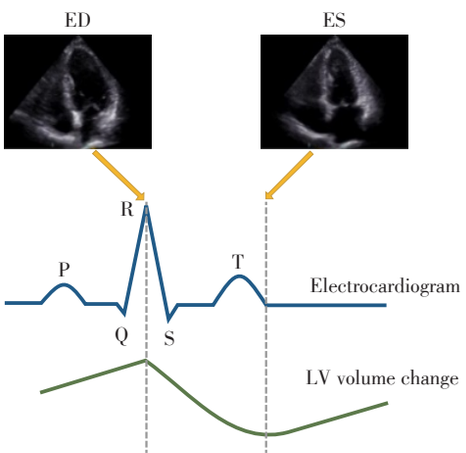


图1 心尖四腔切面ED和ES帧

Figure 1 ED and ES frame of apical four chambers view

针对现阶段相关研究存在的不足，本研究提出

了一种超声心动图关键帧智能检测方法，该方法无需分割LV且不依赖ECG，结合ResNet与Video Swin Transformer(VST)[20]，直接从任意长度的二维心脏超声图像序列中自动、精确地识别出多个心动周期的ED和ES帧。这种方法适用于超声心动图常见的心尖切面，具有较强的实用性。

1　对象和方法

1.1　对象

1.1.1　南京鼓楼医院数据集

选取2022年8—12月在南京大学医学院附属鼓楼医院超声医学科完成二维经胸超声心动图检查的190例受检者，获取包括心尖二腔(apical two chambers，A2C)、心尖三腔(apical three chambers，A3C)与心尖四腔(apical four chambers，A4C)3类临床检查常用切面共计663个动态图像(A2C、A3C与A4C切面的图像数目分别为249、134与280，图2)。上述动态图像均采集自Philips Medical Systems和GE Vingmed Ultrasound设备，帧数在14~493帧之间，且包含不同数量的完整心动周期。研究已获得南京大学医学院附属鼓楼医院医学伦理委员会的伦理审查批准(批件号：2022-337-01)。

为便于图像的后续处理与分析，本研究基于ITK与OpenCV解析DICOM格式的原始超声数据，将其转换为AVI格式的视频。为使数据完全独立于ECG，所得到的视频均基于传统的图像处理技术，包括阈值分割、霍夫直线检测以及形态学操作等进行处理，删除了ECG以及扫描扇区外的文字信息，并使用双线性插值将每帧图像采样到320×320像素。建立的南京鼓楼医院数据集中，每一类切面的数据按4∶1的比例，分别划分到训练集与测试集中。

1.1.2　EchoNet-Dynamic-Tiny数据集

EchoNet-Dynamic数据集[21]源自斯坦福大学医学院2016—2018年收集的受试者数据，包括10 030个A4C视频，采集设备为Siemens Healthineers和Philips Healthcare，每个视频经过匿名化处理，删除了扫描扇区以外的文本等信息，并通过3次降采样插值将每帧图像大小调整为112×112像素。本研究从EchoNet-Dynamic数据集中选取了按视频名称升序排列的前280个A4C视频，与南京鼓楼医院数据集中A4C的数量保持一致，建立了EchoNet-Dynamic-Tiny数据集，并将这些数据按4∶1的比例，分别划分到训练集与测试集中。

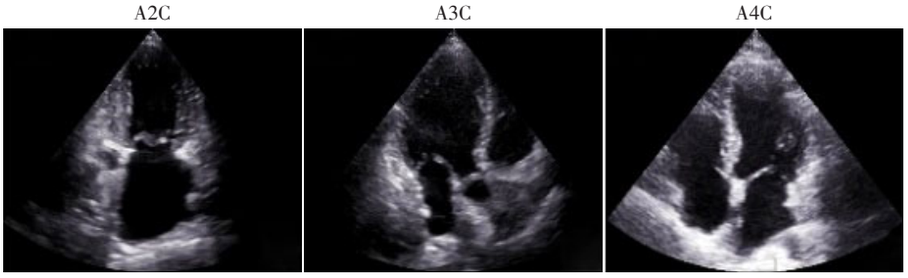


图2　二维超声心动图3类切面

Figure 2　Three types of two-dimensional echocardiography views

1.2　方法

1.2.1　数据标记

在3名经验丰富的超声医师的指导下，对所有视频进行手动标记(对于标记不一致的视频，3位医师重新对其进行标记，然后取多数投票结果)，确定各心动周期的ED、ES帧索引。基于高斯分布计算视频各帧为关键帧的概率，见式1，其中x表示帧索引，μ表示ED或ES帧所在的帧索引，$\rho(x)$表示每一帧为ED或ES帧的概率，σ为常数，本研究中设置为10。

$$\rho(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{式1}$$

将当前心动周期ED帧与下一相邻心动周期ED帧的概率设置为1.0，两帧之间其余帧的概率基于高斯概率密度函数进行插值，得到该心动周期中每一帧为ED帧的概率曲线，ES帧概率曲线的生成方式与之类似。值得注意的是，为确保概率密度函数在视频起始与终止阶段设定合理，本研究对起始与终止阶段的关键帧进行了假定(即μ可能为负数或大于视频总帧数的值)，并在训练与测试过程中仅保留视频总帧数范围内帧索引对应的概率曲线。假设当前视频总帧数为58，标记的ED帧索引为20、66，ES帧索引为-1、43，生成ED、ES概率曲线(图3)。

1.2.2　ResNet+VST模型

本研究提出的超声心动图关键帧智能检测模型ResNet+VST，其整体框架见图4，该模型采用

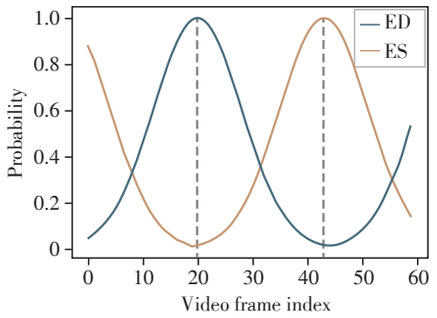
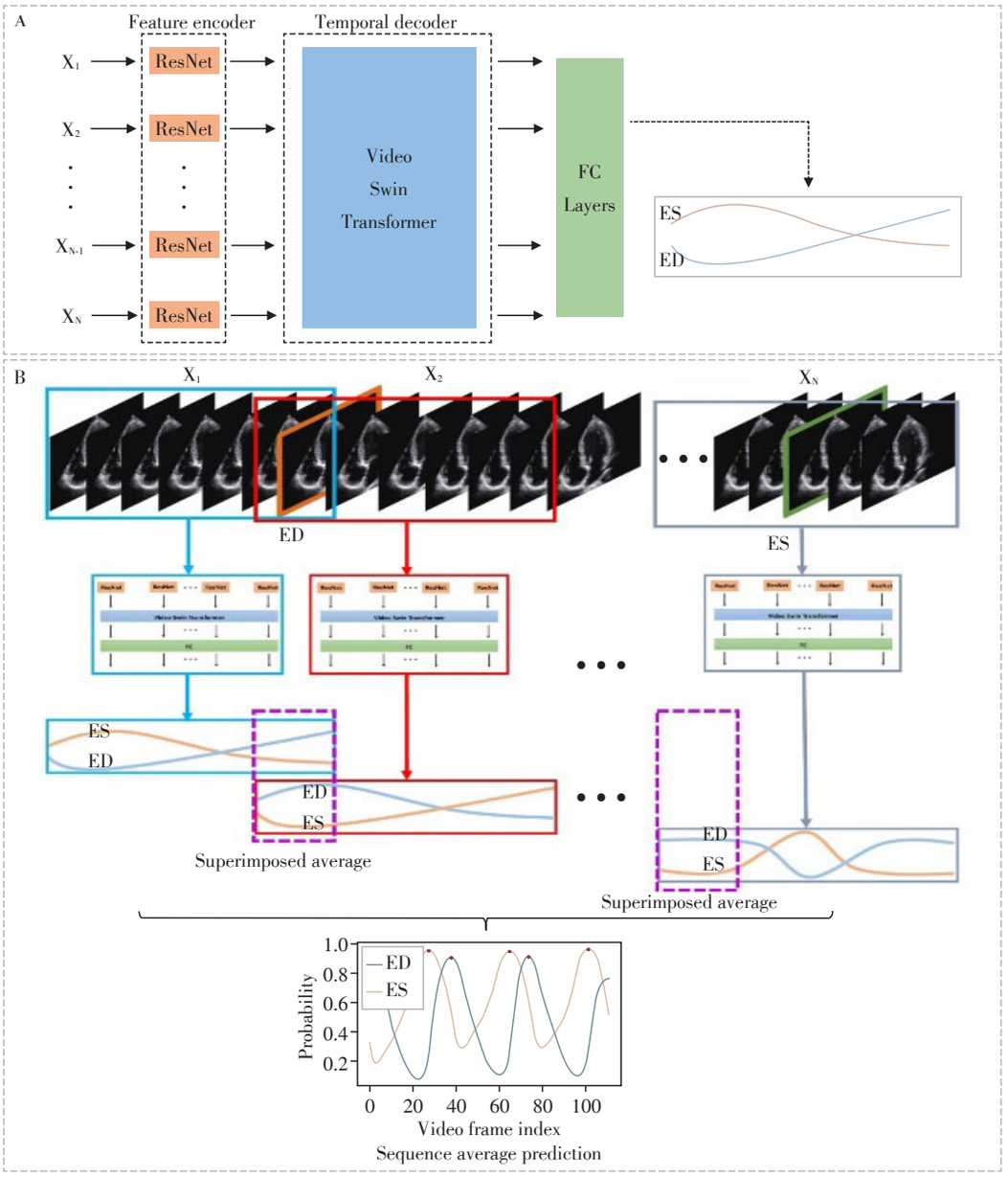


图3 每一帧为ED帧或ES帧的概率

Figure 3 The probability of each frame being an ED frame or and ES frame

CNN与视频旋转变压器VST的混合架构，旨在分析心脏超声视频并提取其时空特征。首先使用CNN捕获输入视频帧的空间特征，然后结合VST提取视频中的时间动态信息，将这些特征整合到全连接层中，最终输出各帧为ED或ES帧的预测概率（图4A）。利用滑动窗口对任意长度超声序列推理（图4B），首先使用滑动窗口对超声视频进行分割，生成固定长度、重叠、分块的超声序列片段，其次将各序列片段输入到神经网络以生成各帧为关键帧的概率，将所有帧在关联窗口下预测值的均值作为最终结果。



A: The neural network architecture. B: The process of inferencing by using sliding window technology.

图4 超声心动图关键帧智能检测方法框架

Figure 4 Framework of intelligent detection method for key frames of echocardiography

神经网络架构：为充分提取超声心动图的时空信息，采用深度神经网络架构ResNet作为编码器，对超声序列中每一帧的空间特征进行编码，然后将获取的空间特征传递给VST，以捕获这些空间特征之间的时间依赖关系。

空间特征提取：首先，使用ResNet编码器从每帧图像中捕获空间特征。ResNet网络由一系列的残差块堆叠而成，每个块包含多个卷积层，通过跨层连接构造本体映射x和残差映射$F(x)$，最终学习的结果为$H(x)=F(x)+x$，这种结构有效地解决了深层网络训练时可能存在的梯度弥散问题，利于网络提取图像深层特征。

时间特征提取：单帧图像空间特征被传递到VST，以进行输入序列片段各帧之间的关联信息提取。VST由模型阶段和头2个部分组成。模型阶段由多个重复的阶段组成，每个阶段包括VST块和融合块。VST模块引入了视频窗口多头自注意力机制(video windows multi-head self-attention，video W-MSA)和视频位移窗口多头自注意力机制(video shifted windows multi-head self-attention，video SW-MSA)，允许在局部窗口内并行计算，以捕获视频序列中的长程时空依赖关系。融合块类似于最大池化，用于降采样、增加通道数，同时保持视频帧数不变。经过模型阶段之后，获得多帧数据的高维特征，最后使用头进行特征融合。完整的VST块结构见图5。

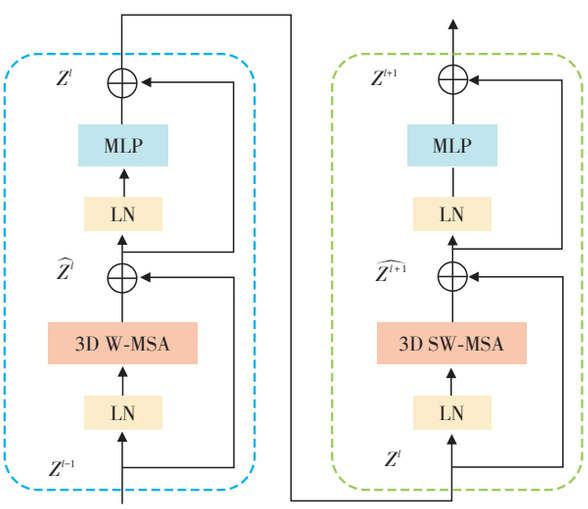


图5　VST模块

Figure 5　VST module

模型推理：为使模型能够处理任意长度的二维超声心动图视频，引入滑动窗口技术，将视频划分为多个重叠的序列片段，输入到神经网络中，以获得各序列片段中每一帧被预测为ED或ES帧的概率值。接着，对原始视频各帧关联窗口的所有预测值求均值，从而得到相应帧为关键帧的概率。计算见式2，其中$\widehat{y_{n,i}}$为第n个序列片段中第i帧的预测值，$\hat{y}_t$为原始视频中第t帧为关键帧的概率值，N代表原始视频每帧关联的窗口数量。最后，通过查找概率的极大值确定网络预测的原始视频关键帧，曲线极大值用红色点表示(图6)。

$$\hat{y}_t=\frac{1}{N}\sum_{n=1}^{N}\widehat{y_{n,\,1}} \qquad 式2$$

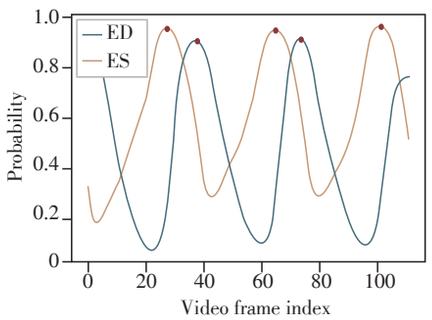


图6　网络预测结果

Figure 6　Network prediction result

1.2.3　实验细节

实验环境：配置见表1。

表1　实验环境配置

Table 1　Configuration of experimental environment

Configuration	Computer
System	Windows 10
Programming language	Python 3.8
DL framework	Pytorch-GPU 1.10
CPU	Intel(R) Core(TM) i5-12600KF
GPU	NVIDIA GeForce ® GTX 3090Ti
RAM	32GB 3200MHz

数据增强：为提高模型的泛化能力并减少过拟合，在模型训练阶段对数据进行了空间和时间两个维度的增强。空间数据增强包括：缩放、随机旋转，值得注意的是，所使用的心脏超声数据MV结构具有特定的方向和生理特征，因此不宜用翻转操作；时间数据增强方面，随机以1、2、4的步长对视频帧进行等间隔采样，以丰富样本的时间尺度，强化模型对不同尺度时间特征的提取能力，若采样至视频末端，序列帧数小于采样大小，则补充零帧。

训练细节：在南京鼓楼医院数据集上，网络输入图像大小为320×320像素；在EchoNet-Dynamic-Tiny数据集上，网络输入图像大小为112×112像

素。使用ResNet50作为空间特征编码器，并使用在ImageNet数据集上经过预训练的模型权重，以提高训练效率。选用Adam网络优化器，初始学习率设置为10^{-5}，并采用多步长学习率衰减策略，训练轮数（epoch）设为500，在现有数据集规模下，批大小（batch size）设置为4，设置序列采样大小为16，以均衡物理显存与时序感受野，丢弃单元（dropout）概率为0.5，以增强模型泛化性，损失函数采用均方误差（mean square error，MSE），以提高对异常值的敏感度并确保关键帧概率值的时间平滑性（式3），其中$Y_{n,t}$和$\widehat{Y_{n,t}}$分别表示为第n个样本中第t帧的真实标签和网络预测概率值。

$$L_{MSE}=\sum_{n=1}^{N}\sum_{t=1}^{T}\left(Y_{n,t}-\widehat{Y_{n,t}}\right)^2 \quad \text{式3}$$

测试细节：在南京鼓楼医院数据集上，网络输入图像大小为224×224像素，在EchoNet-Dynamic-Tiny数据集上，网络输入图像大小为112×112像素。帧采样步长设置为1，滑动窗口步长统一设置为1，窗口宽度统一设置为16帧，在处理每个视频末端时，若序列长度<16帧，则在其末尾填充0帧，且计算关键帧概率时，0帧不纳入计算范畴。

1.2.4 评估指标

使用平均帧差（average frame difference，AFD）[11]衡量所提出方法的预测结果与真实标签之间的绝对误差大小（式4、5），其中，y_i代表ED或ES帧的真实标签，$\widehat{y_i}$表示ED或ES的预测帧索引，N是测试集内ED或ES帧的总数量。

$$\text{AFD}=\frac{1}{N}\sum_{i=1}^{N}\left|y_i-\widehat{y_i}\right| \quad \text{式4}$$

$$\text{std}=sqrt\left(\frac{1}{N}\sum\nolimits_{i=1}^{N}\left|y_i-\widehat{y_i}\right|^2\right) \quad \text{式5}$$

1.3 统计学方法

实验数据采用Python 3.8软件进行统计学分析，计量资料以均数±标准差（$\bar{x}\pm s$）表示，计数资料以百分数（%）表示。多组样本均数比较采用单因素方差分析（one-way ANOVA），$P<0.05$为差异有统计学意义。

2 结 果

为详尽分析ResNet+VST模型的准确性与鲁棒性，本研究分别在南京鼓楼医院数据集与EchoNet-Dynamic-Tiny数据集上进行实验，使用1.2.4节评估指标衡量模型在测试集上的准确性。

本研究在南京鼓楼医院数据集上，分析了ResNet+VST模型预测结果与真实标签的差异，表明其准确性（2.1.1），并将其与现阶段较为先进的3D CNN + LSTM[10]与ResNet + LSTM[11]模型进行对比，表明其先进性（2.1.2）；进一步地，本研究基于公开数据集EchoNet-Dynamic构建的EchoNet-Dynamic-Tiny子数据集上，分析了前述3种模型相应的性能表现（2.2），更充分地衡量ResNet+VST模型的泛化性，便于后续研究者对该模型性能表现进行更客观详尽的评估。

2.1 南京鼓楼医院数据集

2.1.1 模型预测结果与真实标签对比

在A2C、A3C、A4C切面上，ResNet+VST模型的心动周期检出率均高于97%，ED、ES的AFD均小于1.65（表2），且模型预测值与真实标签之间显示出高度一致性（图7）。

在A2C、A3C和A4C切面中各随机挑选1个视频，将ResNet+VST模型的关键帧检测结果与视频帧进行匹配。对于A2C的ED、ES帧，以及A3C的ED帧，预测结果均与人工标注仅相差1帧，且预测帧与真实标签帧的图像内容较为接近（图8A、B）；对于A3C的ES帧与A4C的ED、ES帧，预测结果与人工标注完全一致（图8B、C）。

2.1.2 不同模型对比

在相同的预处理、数据增强和超参数设置下，将所提出的模型与3D CNN+LSTM[10]和ResNet+LSTM[11]关键帧检测模型在临床应用场景更多的A4C切面上进行比较（表3）。ResNet+VST模型在检测精度、推理时间方面均优于其他两个模型。单因素方差分析结果显示，3种模型之间存在显著性差

表2 3类切面模型心动周期检出率与关键帧检测平均帧差情况

Table 2 Detection rate of cardiac cycle and average frame difference of key frame detection in three types of view models

View types	Test videos	Cardiac cycle			AFD($\bar{x}\pm s$)	
		Predicted number	True number	Detection rate(%)	ED	ES
A2C	50	106.0	107.0	99.06	1.52 ± 1.09	1.56 ± 1.16
A3C	27	45.0	45.0	100.00	1.62 ± 1.43	1.63 ± 1.25
A4C	56	114.0	116.5	97.85	1.27 ± 1.17	1.45 ± 1.38

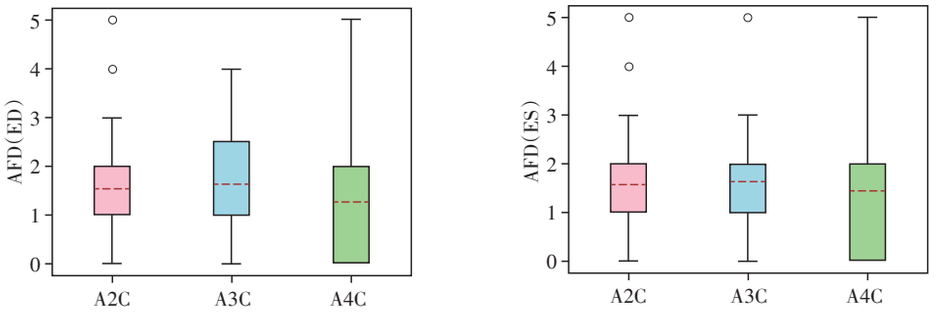
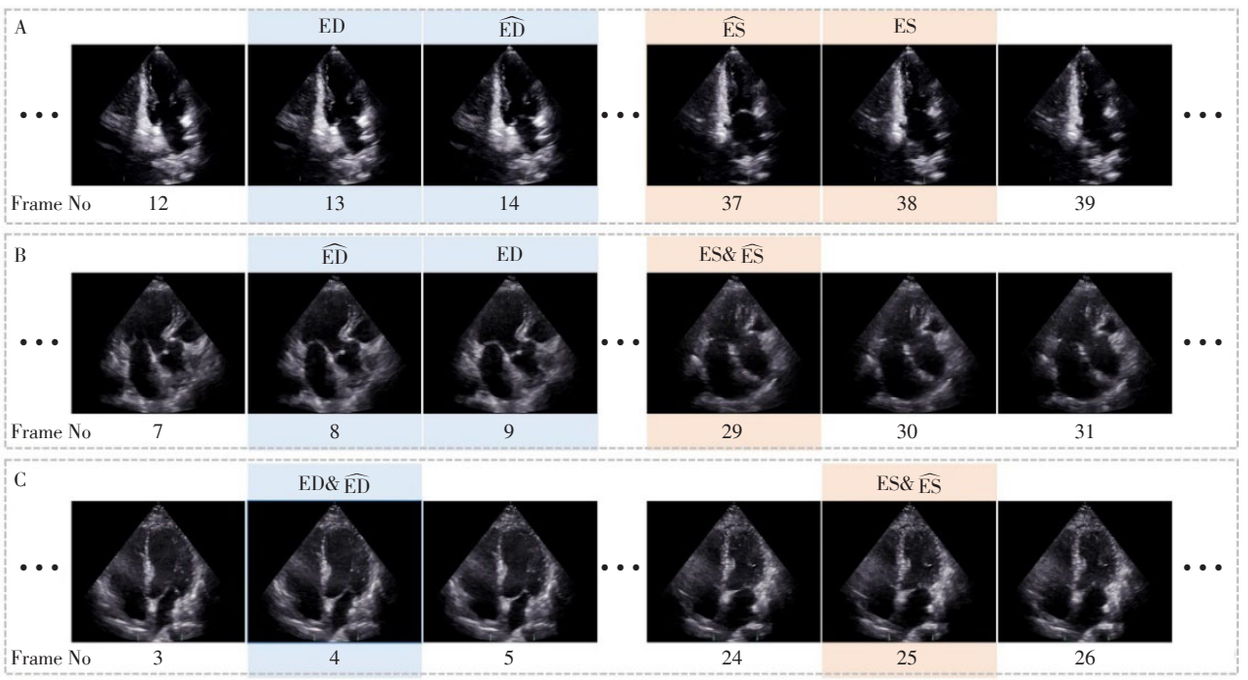


图7　ED和ES模型预测与真实标签一致性对比

Figure 7　Comparison of consistency between ED and ES model prediction and label



A：Model prediction of A2C views and examples of true labels. B：Model prediction of A3C views and examples of true labels. C：Model prediction of A4C views and examples of true labels. ED、ES：true label；$\widehat{ED}$、$\widehat{ES}$：model prediction

图8　A2C、A3C和A4C切面上ResNet+VST模型关键帧检测结果与真实标签对应视频帧示例

Figure 8　Examples of video frames corresponding to the detection results of key frames of ResNet+VST model and labels on A2C，A3C and A4C views

异（$P<0.05$）。Tukey检验结果进一步证明，ResNet+VST模型与3D CNN+LSTM以及ResNet+LSTM模型之间均存在显著性差异（$P<0.05$）。

2.2　EchoNet-Dynamic-Tiny数据集

从EchoNet-Dynamic-Tiny数据集中随机挑选1个视频，将ResNet+VST模型的关键帧检测结果与视

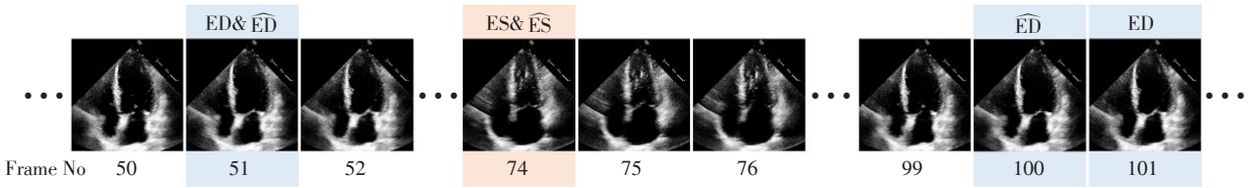
表3　南京鼓楼医院数据集A4C切面不同模型ED、ES帧检测误差与推理时间对比

Table 3　Comparison of detection error and inferencing time of ED and ES frames of different models on A4C view of Nanjing Drum Tower Hospital dataset　($\bar{x} \pm s$)

Performance	Model			F	P
	A	B	C		
AFD(ED)	1.270 ± 1.170*#	1.810 ± 1.690#	1.900 ± 1.880	5.535	0.004
AFD(ES)	1.450 ± 1.380*#	1.920 ± 1.850#	1.650 ± 1.560	3.591	0.028
Inference time(s)	0.021 ± 0.002*#	0.157 ± 0.009#	0.136 ± 0.005	>100.000	<0.001

A：ResNet+VST model. B：3D CNN+LSTM model. C：ResNet+LSTM model. Compared with B，*$P<0.05$；Compared with C，#$P<0.05$.

频帧进行匹配。结果显示模型的ES预测帧与真实标签完全一致，而ED预测帧与真实标签帧非常接近且图像内容相似度较高（图9）。



ED, ES: true label; $\widehat{ED}$, $\widehat{ES}$: model prediction.

图9 EchoNet-Dynamic-Tiny数据集ResNet+VST模型关键帧检测结果与真实标签对应视频帧示例

Figure 9 Example of video frames corresponding to the keyframe detection results of the EchoNet-Dynamic-Tiny dataset ResNet+VST model and label

动图数据集，所提出的ResNet+VST模型在关键帧检测任务中的预测结果与真实标签更为接近，表现出更高的准确性与更快的推理速度。单因素方差分析结果显示，3种模型之间存在显著性差异（$P < 0.05$）。Tukey检验结果进一步证明，ResNet+VST模型与3D CNN+LSTM以及ResNet+LSTM模型之间均存在显著性差异（$P < 0.05$）。

使用EchoNet-Dynamic-Tiny数据集，将所提出的ResNet+VST模型与3D CNN+LSTM[10]、ResNet+LSTM[11]模型进行比较（表4）。针对公开的超声心

3 讨 论

超声心动图凭借无创、无辐射、安全等特点，成为心脏疾病诊断的主要医学影像手段。其中，ED和ES帧检测对于评估超声心动图图像质量和测量心脏参数至关重要。临床上ED和ES帧的选定主要依靠医师借助ECG或目测LV的容积，可重复性差，相

表4 EchoNet-Dynamic-Tiny数据集A4C切面不同模型ED、ES帧检测误差与推理时间对比

Table 4 Comparison of detection error and inferencing time of ED and ES frames of different models on A4C view of EchoNet-Dynamic-Tiny dataset $(\bar{x} \pm s)$

Performance	Model			F	P
	A	B	C		
AFD(ED)	1.620 ± 1.260*#	1.770 ± 1.470#	1.830 ± 1.680	2.892	0.037
AFD(ES)	1.710 ± 1.180*#	1.980 ± 1.660#	1.810 ± 1.750	3.026	0.032
Inference time(s)	0.010 ± 0.001*#	0.141 ± 0.003#	0.120 ± 0.001	>100	<0.001

A: ResNet+VST model. B: 3D CNN+LSTM model. C: ResNet+LSTM model. Compared with B, *$P < 0.05$; Compared with C, #$P < 0.05$.

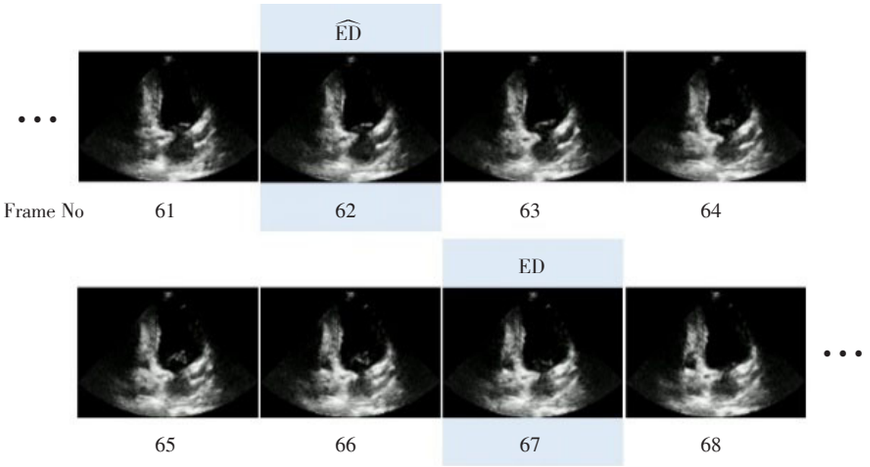
比之下，自动检测快速、高效、可重复性好。目前已有一些基于DL的超声心动图关键帧智能检测方法[6-7,11-14,16-18]，但它们主要关注A4C切面，并且无法同时满足检测精度和推理耗时的要求。为解决这些问题，本研究提出了一种新的关键帧检测模型ResNet+VST，该模型结合了带有跨层连接的ResNet和带有自注意力机制的VST，能够有效提取超声序列图像的复杂时空信息，并结合曲线回归策略，将网络输出回归为关键帧的概率，将复杂的关键帧检测问题转化为概率曲线回归问题。

本研究结果表明，ResNet+VST模型在南京鼓楼医院数据集和EchoNet-Dynamic-Tiny数据集上表现良好。在南京鼓楼医院数据集A2C、A3C、A4C 3类切面上，模型预测的心动周期数量与真实数量均较为接近，证明了所提出方法的有效性，模型预测的ED和ES帧与真实标签之间的AFD均小于1.65，表明模型拥有较高的准确率且预测值与真实标签之间显示出高度一致性；在EchoNet-Dynamic-Tiny数据集A4C切面上，模型预测的ED和ES帧的AFD均小于1.75，且相比3D CNN+LSTM[10]、ResNet+LSTM[11]模型，ResNet+VST模型在2个数据集上的预测结果更接近真实标签，各模型预测结果之间均存在显著性差异。此外，与利用LSTM进行时序建模的相关研究[10-11]相比，ResNet+VST模型计算上高度并行，减少了推理时间的消耗，在Intel（R）Core（TM）i5-12600KF CPU与NVIDIA GeForce® GTX 3090Ti GPU

的硬件条件下，在南京鼓楼医院数据集上，当输入图像大小设置为224×224像素时，16帧的超声序列片段推理平均耗时仅为21 ms，而在EchoNet-Dynamic-Tiny数据集上，当图像大小设置为112×112像素时，推理耗时更短，仅为10 ms，基本满足临床需求。

然而，本研究仍存在不足之处：①数据量较少，ResNet+VST模型检测性能的更全面、充分评估，需要更大规模的数据集支持；②数据来源较为单一，仅在两个数据集上进行实验，对模型泛化性的评估能效不足；③部分数据中关键帧位置位于超声动态图像的起止端，ResNet+VST模型在这类问题上表现欠佳，后续考虑结合时序建模领域的最新研究，如新型RNN架构[22]；④部分样本的射血阶段持续时间较短，与ED之间的时间间隔较小，这容易导致标注者之间、标注者内部的标注结果存在分歧，自然地，模型同样可能将射血阶段帧误判为ED帧，与真实标签存在较大偏差(图10)。因此，进一步研究标注者之间和标注者内部的差异是必要的。

综上所述，本研究所提出的基于DL的超声心动图关键帧智能检测模型ResNet+VST，适用任意长



ED：true label；$\widehat{\text{ED}}$：model prediction.

图10　ResNet+VST模型预测异常示例

Figure 10　Example of abnormal prediction by ResNet+VST model

度且包含多个心动周期不同切面的超声心动图。与目前常用的方法相比，该模型在检测精度和速度方面均具有显著的优势，基本满足临床及实际应用需求，具有良好的应用价值。未来该模型有望作为超声心动图图像质量自动评估以及腔室容积、射血分数等心脏参数自动测量的预处理步骤，以实现更准确和快速的心脏图像分析。

[参考文献]

[1] VIEILLARD-BARON A，MILLINGTON S J，SANFILIPPO F，et al. A decade of progress in critical care echocardiography：a narrative review[J]. Intensive Care Med，2019，45(6)：770-788

[2] KHEIWA A，HARRIS I S，VARADARAJAN P. A practical guide to echocardiographic evaluation of adult Fontan patients[J]. Echocardiography，2020，37(12)：2222-2230

[3] SIRJANI N，MORADI S，OGHLI M G，et al. Automatic cardiac evaluations using a deep video object segmentation network[J]. Insights Imag，2022，13(1)：69

[4] 张　钒，陆正大，李春迎，等. 超声图像分割的研究进展[J]. 临床超声医学杂志，2022，24(6)：453-456

[5] 蒋建慧，姚　静，张艳娟，等. 基于深度学习的超声自动测量左室射血分数的研究[J]. 临床超声医学杂志，2019，21(1)：70-74

[6] ZENG Y，TSUI P H，PANG K J，et al. MAEF-Net：multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography[J]. Ultrasonics，2023，127：106855

[7] DARVISHI S，BEHNAM H，POULADIAN M，et al. Measuring left ventricular volumes in two-dimensional echocardiography image sequence using level-set method for automatic detection of end-diastole and end-systole frames[J]. Res Cardiovasc Med，2013，2(1)：39-45

[8] KUSUNOSE K. Radiomics in echocardiography：deep learning and echocardiographic analysis[J]. Curr Cardiol Rep，2020，22(9)：89

[9] 罗　刚，洋思林，乔思波，等. 深度学习技术在胎儿超声心动图图像自动识别中的应用[J]. 实用医学杂志，2022，38(14)：1830-1833

[10] 吴 洋,张红梅,尹立雪,等.超声心动图心尖四腔心切面图像质量智能评分研究[J].中华医学超声杂志(电子版),2023,20(1):97-102

[11] 李 敬,刘宁宁,王笑一.人工智能在超声诊断中的应用现状及展望[J].中国超声医学杂志,2022,38(5):595-598

[12] FARHAD M,MASUD M M,BEG A. Deep learning based cardiac phase detection using echocardiography imaging [C]//International Conference on Advanced Data Mining and Applications.Cham:Springer,2022:3-17

[13] DEZAKI F T,DHUNGEL N,ABDI A H,et al. Deep residual recurrent neural networks for characterisation of cardiac cycle phase from echocardiograms [C]//International Workshop on Deep Learning in Medical Image Analysis, International Workshop on Multimodal Learning for Clinical Decision Support.Cham:Springer,2017:100-108

[14] KONG B,ZHAN Y Q,SHIN M,et al. Recognizing end-Diastole and end-Systole frames via deep temporal regression network [M]//OURSELIN S, JOSKOWICZ L, SABUNCU M R, et al., Eds.Lecture Notes in Computer Science.Cham: Springer International Publishing, 2016: 264-272

[15] HE F X,LIU T L,TAO D C. Why ResNet works?Residuals generalize[J]. IEEE Trans Neural Netw Learn Syst, 2020,31(12):5349-5362

[16] QIN C,CHEN L M,CAI Z T,et al. Long short-term memory with activation on gradient [J]. Neural Netw, 2023, 164:135-145

[17] TAHERI DEZAKI F,LIAO Z B,LUONG C,et al. Cardiac phase detection in echocardiograms with densely gated recurrent neural networks and global extrema loss[J]. IEEE Trans Med Imag,2018,38(8):1821-1832

[18] FIORITO A M,ØSTVIK A,SMISTAD E,et al. Detection of cardiac events in echocardiography using 3D convolutional recurrent neural networks[C]//2018 IEEE International Ultrasonics Symposium (IUS).Kobe, Japan.IEEE, 2019:1-4

[19] LANE E S,AZARMEHR N,JEVSIKOV J,et al. Multibeat echocardiographic phase detection using deep neural networks[J]. Comput Biol Med,2021,133:104373

[20] LIU Z,NING J,CAO Y,et al. Video swin transformer[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). New Orleans,LA,USA.IEEE, 2022:3192-3201

[21] OUYANG D,HE B,GHORBANI A,et al. Video-based AI for beat-to-beat assessment of cardiac function[J]. Nature,2020,580(7802):252-256

[22] YIN H,ZHOU Y H,CAO L,et al. Channel prediction with liquid time-constant networks:an online and adaptive approach[C]//2021 IEEE 94th Vehicular Technology Conference(VTC2021-Fall). Norman,OK,USA.IEEE, 2021:1-6

[收稿日期] 2023-08-08

(本文编辑:戴王娟)