E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)01-0268-14

论文引用格式:Li Y C, Chen D L, Guo D H and Sun Y. 2025. Integrating spatiotemporal features and temporal constraints for dual-modal breast tumor diagnosis. Journal of Image and Graphics, 30(01):0268-0281(李一宸, 陈大力, 郭丁豪, 孙羽. 2025. 融合时空特征与时间约束的双模态乳腺肿瘤诊断. 中国图象图形学报, 30(01):0268-0281)[DOI:10.11834/jig.240217]

融合时空特征与时间约束的双模态乳腺肿瘤诊断

李一宸,陈大力*,郭丁豪,孙羽

东北大学信息科学与工程学院,沈阳　110819

摘　要:**目的**　综合考虑B型超声(B-mode ultrasound, B-US)和对比增强超声(contrast-enhanced ultrasound, CEUS)双模态信息有助于提升乳腺肿瘤诊断的准确性,从而利于提高患者生存率。然而,目前大多数模型只关注B-US的特征提取,忽视了CEUS特征的学习和双模态信息的融合处理。为解决上述问题,提出了一个融合时空特征与时间约束的双模态乳腺肿瘤诊断模型(spatio-temporal feature and temporal-constrained model, STFTCM)。**方法**　首先,基于双模态信息的数据特点,采用异构双分支网络学习B-US和CEUS包含的时空特征。然后,设计时间注意力损失函数引导CEUS分支关注造影剂流入病灶区的时间窗口,从该窗口期内提取CEUS特征。最后,借助特征融合模块实现双分支网络之间的横向连接,通过将B-US特征作为CEUS分支补充信息的方式,完成双模态特征融合。**结果**　在收集到的数据集上进行对比实验,STFTCM预测的正确率、敏感性、宏平均F1和AUC(area under the curve)指标均表现优秀,其中预测正确率达88.2%,领先于其他先进模型。消融实验中,时间注意力约束将模型预测正确率提升5.8%,特征融合使得模型诊断正确率相较于单分支模型至少提升2.9%。**结论**　本文提出的STFTCM能有效地提取并融合处理B-US和CEUS双模态信息,给出准确的诊断结果。同时,时间注意力约束和特征融合模块可以显著地提升模型性能。

关键词:双模态乳腺肿瘤诊断;时空特征;时间注意力约束;对比增强超声(CEUS);B型超声(B-US)

Integrating spatiotemporal features and temporal constraints for dual-modal breast tumor diagnosis

Li Yichen, Chen Dali*, Guo Dinghao, Sun Yu

College of Information Science and Engineering, Northeastern University, Shenyang　110819, China

Abstract: **Objective** Breast cancer ranks first in the incidence of cancer among women worldwide, impacting the health of the female population. Timely diagnosis of breast tumor can offer better treatment opportunities for patients. B-mode ultrasound (B-US) imaging contains rich spatial information such as lesion size and morphology. It is widely used in breast tumor diagnosis because of its advantages of low cost and high safety. On this basis, with the advancement of deep learning technology, some deep learning models have been applied to computer-aided diagnosis of breast tumor diagnosis based on B-US to assist doctors. However, diagnosis based solely on B-US imaging results in lower specificity. Moreover, the perfor-

收稿日期:2024-04-12;**修回日期**:2024-05-13;**预印本日期**:2024-05-20

*通信作者:陈大力　chendali@ise.neu.edu.cn

基金项目:国家自然科学基金项目(61773104);国家级大学生创新创业训练计划项目(202310145018);中央高校基本科研业务专项资金资助(N2404011)

Supported by: National Natural Science Foundation of China (61773104); National Training Program of Innovation and Entrepreneurship for Undergraduates (202310145018); Fundamental Research Funds for the Central Universities (N2404011)

mance of models trained exclusively on B-US is limited by the singular modality of information source. Contrast-enhanced ultrasound (CEUS) can provide a second modality of information on top of B-US to improve diagnostic accuracy. CEUS contains rich spatiotemporal information, such as brightness enhancement and vascular distribution in the lesion area, by injecting contrast agents intravenously and capturing the information during the time window when the contrast agent flows into the lesion area. Considering the B-US and CEUS dual-mode information comprehensively can enhance diagnostic accuracy. A model integrating spatiotemporal features and temporal-constrained (STFTCM) for dual-modality breast tumor diagnosis is proposed to effectively utilize dual-modal data for breast tumor diagnosis. **Method** STFTCM primarily comprises a heterogeneous dual-branch network, a temporal attention constraint module, and feature fusion modules. On the basis of the characteristics of dual-mode data information dimensions, STFTCM adopts a heterogeneous dual-branch structure to extract the dual-mode feature separately. For the B-US branch, B-US consists of spatial features within the two-dimensional frames of the video, and inter-frame transformations are not prominent. Considering that training 3D convolutional networks on a small dataset tends to result in overfitting due to the larger number of parameters compared with 2D convolutional networks, a 2D network, ResNet-18, is used as the backbone network for feature extraction from a single frame extracted from the video. In contrast, CEUS video frames undergo noticeable transformations during the time window when the contrast agent flows through the lesion area, containing rich spatiotemporal information. Thus, a 3D network, R(2 + 1)D-18, is used as the backbone network for the CEUS branch. The structure is adjusted on the basis of the backbone network to ensure the feature maps extracted from corresponding layers of the dual-branch network have the same dimensions for subsequent dual-mode fusion. On the basis of the aforementioned CEUS branch, the spatiotemporal information in CEUS mainly resides within the time window when the contrast agent flows into the lesion area; thus, guiding the model to focus on this time segment facilitates better learning of CEUS features on a small dataset. To address this issue, a temporal attention loss function is proposed. An analysis of the temporal knowledge of CEUS video shows that the temporal loss function first determines the temporal attention boundary on the basis of the first-order difference of the discrete sequence of CEUS frames luminance values and then establishes a temporal mask. Subsequently, the temporal attention loss function guides the updating of the parameters of the R(2 + 1)D temporal convolutional kernels in the CEUS branch based on the temporal mask, thereby directing the model to focus on the information during the period when the contrast agent flows into the lesion area. Furthermore, a feature fusion module is introduced to fuse dual-mode information and thus improve prediction accuracy by considering information from both B-US and CEUS. Model parameter size is controlled by not setting a separate third feature fusion branch; instead, the feature fusion module facilitates feature fusion between the dual branches through lateral connections. B-US spatial information is incorporated as supplementary data into the CEUS branch. The feature fusion module comprises a spatial feature fusion module and an identity mapping branch. The spatial feature fusion module combines two-dimensional spatial feature maps of B-US and three-dimensional spatiotemporal feature maps of CEUS, while the identity mapping branch prevents loss of the original CEUS features during the fusion process. **Result** Comparative and structural ablation experiments are conducted to explore the performance of STFTCM and the effectiveness of the temporal attention constraint and feature fusion modules. Experimental data are obtained from the Shengjing Hospital of China Medical University, comprising 332 ultrasound contrast videos categorized into benign tumors, malignant tumors, and inflammations, with 101, 102, and 129 instances, respectively. Accuracy, sensitivity, specificity, macro-average F1, and area under the curve are used as model performance evaluation metrics. In comparative experiments, STFTCM achieves an accuracy of 0.882, with respective scores of 0.909, 0.870, 0.883, and 0.952 for the other four metrics, outperforming other advanced models. In single-branch model comparison experiments, both the B-US and CEUS branches of STFTCM perform better than other advanced models. Comparative experiments between dual-branch and single-branch models demonstrate the excellent performance of STFTCM. Structural ablation experiment results show that temporal attention loss constraint improved prediction accuracy by 5.8 percentage points, and dual-modal feature fusion enhanced prediction accuracy by at least 2.9 percentage points compared with unimodal predictions, confirming the effectiveness of the temporal attention constraint and feature fusion modules in enhancing model performance. Additionally, visualization of model attention through class activation maps validates that the temporal attention constraint improves the model's attention in the temporal dimension, guiding better extraction of spatiotemporal information contained in CEUS. Results from experiments

related to the feature fusion module demonstrate that the addition of the identity mapping branch could improve prediction accuracy by 2.9 percentage points, further confirming the rationality of the feature fusion module's structural design. **Conclusion** STFTCM, designed based on prior knowledge, has demonstrated excellent performance for breast tumor diagnosis. The heterogeneous dual-branch structure designed on the basis of the characteristics of dual-mode data effectively extracts B-US and CEUS dual-mode features while reducing model parameters for better parameter optimization on small datasets. The temporal attention loss function constrains the model's attention in the temporal dimension, guiding the model to focus on information from the time window when the contrast agent flows into the lesion area. Furthermore, the feature fusion module effectively implements lateral connections between dual-branch networks to fuse dual-mode features.

Key words: dual-modality breast tumor diagnosis; spatiotemporal feature; temporal attention constraint; contrast-enhanced ultrasound(CEUS); B-mode ultrasound(B-US)

0 引 言

乳腺癌在全球女性癌症中发病率居首位(Siegel等,2024),及早诊断可降低乳腺癌患者的死亡率(Anderson等,2003)。B型超声(B-mode ultrasound,B-US)技术因其成本低廉、安全性高等优点(Evans等,2018),广泛应用于乳腺肿瘤诊断(Hooley等,2013)。但是仅依靠B-US进行诊断会导致预测结果特异性较低(Wubulihasimu等,2018)。对此,联合使用B-US和对比增强超声(contrast-enhanced ultrasound, CEUS)技术可以弥补上述不足(Du等,2012)。CEUS通过向患者体内注射相关的超声造影剂获取病灶区相关信息(Guo等,2018),在B-US的基础上为乳腺肿瘤诊断提供额外的依据。

然而,放射科医生基于超声影像做出的诊断结果很大程度上取决于医生本人知识水平和临床经验,不同医生之间存在较大差异;同时,疲劳等因素也会影响医生对肿瘤类型的判断。为了辅助医生提高诊断的准确性和效率,一些基于深度学习模型的计算机辅助诊断系统被提出(Masud等,2019)。但是,目前基于超声影像的乳腺肿瘤诊断模型主要集中于对B-US特征的学习(Luo等,2024),忽视了CEUS特征提取以及双模态特征的融合处理。尤其是在数据不足情况下,如何提取CEUS中病灶区时空维度信息以及如何融合B-US和CEUS两种不同维度的特征,仍是需要解决的问题。

针对上述问题,本文结合先验知识提出了一个融合时空特征与时间约束的双模态乳腺肿瘤诊断模型(spatio-temporal feature and temporal-constrained model, STFTCM)。根据B-US和CEUS数据的时序特点,该模型采用异构双分支网络,使用结构调整后的ResNet-18(residual network)和R(2+1)D-18模型分别提取B-US空间维度特征和CEUS时间维度与空间维度特征。由于目前并未有同时包含B-US和CEUS的较大的公共数据集,为了在收集到的数据集上有目的地且高效地获取CEUS信息,在模型训练过程中,本文设计了时间注意力损失函数引导CEUS分支重点关注造影剂流入病灶区的时间窗口。此外,STFTCM借助特征融合模块,通过双分支网络间横向连接的方式将B-US特征融合进CEUS分支中,实现双模态特征融合。

在收集到的数据集上得到的实验结果表明,STFTCM在肿瘤预测结果正确率等评价指标上优于当前其他先进模型。此外,消融实验等一系列实验结果表明,基于先验知识的时间注意力约束和特征融合模块可以显著地提升模型性能。

本文主要贡献如下:1)结合相关先验知识和数据特点,提出了一个融合时空特征与时间约束的双模态乳腺肿瘤诊断模型;2)基于CEUS数据时序分析,设计了一个时间注意力损失函数,指导模型关注造影剂流入病灶区的时间窗口,学习CEUS中时空信息;3)设计了一个特征融合模块,实现双分支网络间的横向连接。通过将B-US信息作为CEUS分支补充信息的方式进行双模态信息的融合。

1 相关工作

随着深度学习技术的发展,一些模型被用于乳腺肿瘤超声影像的计算机辅助检测系统中(Ilesanmi等,2021)。Byra(2018)使用VGG19(Visual Geometry Group)提取特征并且通过Fisher判别分析(Belhumeur等,1997)评估提取得到的特征进行乳腺肿瘤的诊断。Moon等人(2020)分别用3种卷积神经网

络针对原始超声图像、分割肿瘤图像、肿瘤形状图像和融合图像进行特征提取，通过集成方法综合考虑多模态特征并进行肿瘤分类。然而，上述方法均需要医生手工标注感兴趣的病灶区域。为了减少模型对人工标注的依赖，Lee 等人（2021）使用 Mask R-CNN（mask region-based convolutional neural network）进行肿瘤区域的定位，并在此区域基础上进行特征提取。随着深度学习技术不断完善，一些直接针对原始 B-US 图像而非仅病灶区范围图像的模型被提出。Lu 等人（2022）通过向 ResNet-18（residual network-18）引入空间注意力机制的方式引导模型在空间维度上自适应地学习肿瘤特征。Mo 等人（2023）结合相关先验知识，提出了 HoVer-Trans 块，引导模型在无人工标注感兴趣区域的条件下仍能在空间维度上关注病灶区信息，增强了模型可解释性。此外，贡荣麟等人（2022）采用迁移学习思想，通过两级迁移学习，使得模型可以从 B-US 图像中同时学习到 B-US 和超声弹性成像特征信息。然而上述模型只是针对 B-US 数据进行处理，数据种类的单一性限制了模型性能。

随着造影剂的更新迭代，CEUS 在乳腺肿瘤诊断中显现出独特的优势。同时，近年来针对视频数据的深度学习技术发展迅猛（Zhu 等，2020），一些基于 B-US 和 CEUS 视频数据的双模态乳腺肿瘤诊断模型被提出（Singh 等，2020）。杨子奇等人（2020）基于双流结构针对 B-US 和 CEUS 混合数据提出协同约束网络。为了使得模型更有效地提取病灶区信息，Chen 等人（2021）基于放射科医生先验知识提出了以 3D 卷积网络为主干的乳腺肿瘤诊断模型。赵绪等人（2022）使用双流网络进行乳腺肿瘤分类，通过向 B-US 分支引入空间注意力机制方式提高模型性能。Gong 等人（2023）进一步挖掘放射科先验知识，针对 CEUS 的时间强度曲线和非影像临床数据进行特征的提取，并作为诊断参考依据，同时应用对抗性自适应方法融合多模态特征。在具体模型的基础之上，Guo 等人（2024）提出了一种知识增强框架，将先验知识融合到双流网络中，该框架具有一定的普适性。

然而，上述工作并未引导模型关注 CEUS 视频亮度变换过程中蕴含的信息。对此，本文提出 STFTCM 模型，该模型采用异构双分支网络作为基本框架，从 B-US 和 CEUS 视频中提取双模态特征。在此基础上，设计了一个时间注意力损失函数，以引导模型高效地提取 CEUS 视频中反映的肿瘤信息。此外，本文设计了一个专门用于将二维空间特征与三维时空特征相结合的融合模块，以实现双分支网络间的横向连接，从而完成双模态特征融合。

2 方　法

本节先对整个模型结构进行概述，然后阐述模型中的细节部分。

2.1 概述

STFTCM 结构如图 1 所示，由异构双分支网络、时间注意力模块和特征融合模块组成。基于双模态数据特点，异构双分支网络针对单帧 B-US 图像和多

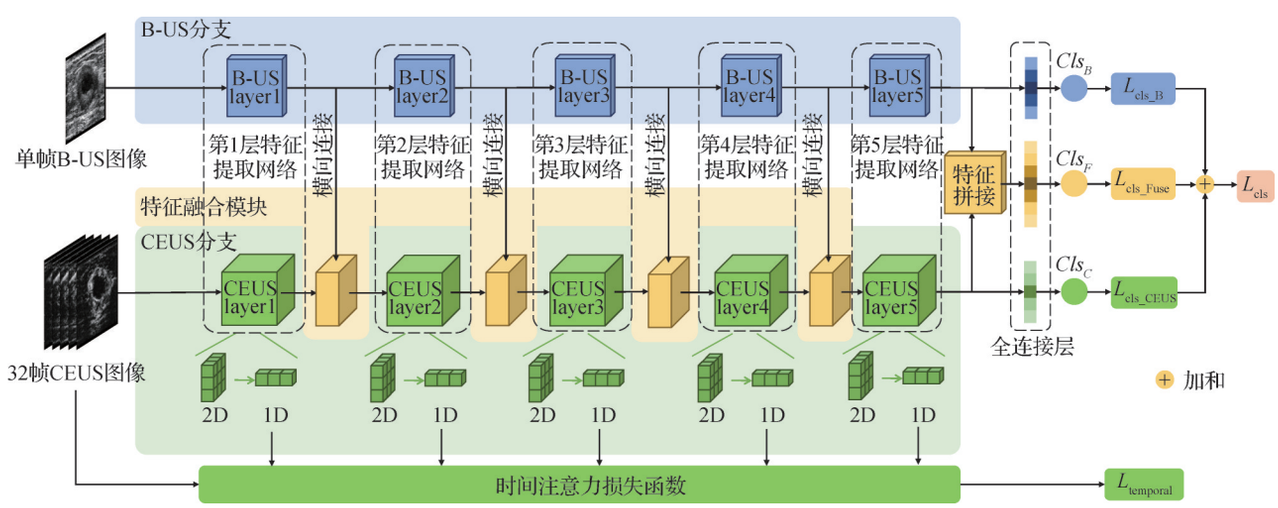


图 1　STFTCM 模型结构

Fig. 1　Structure of STFTCM

帧CEUS图像分别提取空间特征和时空特征。针对双分支网络中的CEUS分支，引入时间注意力模块，引导模型关注特定时间窗口信息，使得CEUS分支模型有效地学习病灶区时空特征。在此基础上，特征融合模块完成维度不对等特征的融合处理，实现双分支网络间的横向连接，在不引入特征融合分支增加模型参数的前提下完成双模态特征的融合。

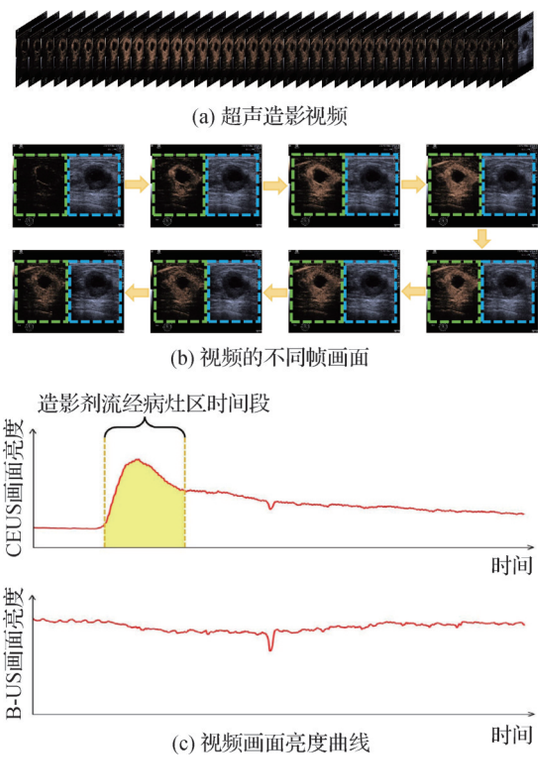
在模型训练阶段，为了使模型更好地提取特征，基于B-US特征、CEUS特征和融合特征的预测结果以及对应的分类损失函数均应考虑。在模型推理阶段，基于融合特征的预测结果作为最终的诊断结果。此外，在模型训练过程中，B-US分支采用了Guo等人(2024)提出的空间损失函数。但是考虑到均方误差的鲁棒性较低，将其替换为平滑L1范数损失函数(Girshick，2015)。

2.2 异构双分支网络

较小的数据集对于模型的训练提出了极大的挑战，特别是对超声影像数据中丰富病理信息的学习，数据量不足容易使得模型过拟合。为了使模型在数据集有限的情况下可以有效地提取B-US和CEUS中丰富的时空特征，双分支主干网络的设计需要兼顾模型的特征提取能力和轻量化。此外，为了便于后续双模态特征融合，异构双分支网络对应层提取到的特征应在空间维度上保持一致。对此本文以双模态数据特点为切入点，针对不同维度数据，提出了异构双分支网络，提取病灶区时空特征，有效地平衡了模型参数量和模型学习能力之间的关系。为了满足后续特征融合的数据维度要求，对主干网络结构进行了相应的调整。

2.2.1 数据特点分析

乳腺超声造影数据如图2所示。图2(b)中，每一帧视频画面的左半部分绿框标出的为CEUS画面，右半部分蓝框标出的为B-US画面。视频画面亮度曲线如图2(c)所示，B-US视频亮度随时间基本保持不变，其二维画面中含有丰富的空间特征，如病灶形状和区域大小等静态信息(Zonderland，2000)。相较于B-US，CEUS视频亮度随时间变换较为明显，病灶区亮度随造影剂的流入和流出呈现出先升后降的趋势。造影剂流经病灶区的时间段内蕴含的时间维度信息，如亮度强化程度、亮度峰值等，有助于区分良性肿瘤和恶性肿瘤。同时，CEUS视频的亮度变换动态过程还反映了病灶区的血管分布和病灶区大小等空间信息(Balleyguier等，2009)。异构双分支网络的设计基于上述数据特点分析，保障模型对于特征学习能力的同时减少模型参数量，缓解训练过程中模型过拟合问题。



(a) 超声造影视频

(b) 视频的不同帧画面

(c) 视频画面亮度曲线

图2 B-US和CEUS视频样例

Fig. 2 B-US and CEUS video sample ((a) ultrasound video sample; (b) different frames of the video; (c) video brightness curves)

2.2.2 B-US分支

3D卷积网络在数据量不足的情况下训练容易过拟合。虽然B-US为视频数据，但是其连续帧之间的画面变化很小，且由数据特点分析可知B-US只含有病灶区的二维空间信息。为了减少模型参数量，B-US分支主干网络选用ResNet-18(He等，2016)，通过二维网络来学习B-US单帧画面中的空间信息。在此基础上，B-US分支引入通道注意力机制(Hu等，2018)以适应不同的通道信息，增强学习能力。

2.2.3 CEUS分支

CEUS特征提取则是基于视频中多帧画面进行的。由于CEUS视频数据的信息量较大，需要CEUS分支模型具有较强的学习能力。此外，由于数据集较小，模型参数需要便于优化。R(2+1)D网络(Tran等，2018)通过将3D卷积核拆散成两个子卷积核的方式增加了模型的非线性，提高了模型的学习

能力，同时，卷积核的拆分有利于模型的优化（Qiu等，2017）。因此，选用R（2 + 1）D-18作为CEUS特征提取的主干网络。

2.2.4　模型结构调整

为了便于后续的特征融合，双分支网络对应层提取的特征应具有相同的维度（C，H，W），因此需要在上述双分支网络的基础上调整模型结构。鉴于3D模型具有较大的参数量，其参数优化过程相比于2D模型更为复杂，结构调整会对3D模型性能产生较大影响。因此，本文选择对B-US分支的模型结构进行调整。修改后的双分支模型结构如表1所示。

表1　双分支结构

Table 1　Dual-branch structure

模型层数	B-US分支	CEUS分支
第1层	7 × 7, 64, 步长为2 SE_block1	3 × 7 × 7, 64, 步长为1 × 2 × 2
第2层	$\begin{pmatrix}3\times3,64\\3\times3,64\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\times3,64\\3\times3\times3,64\end{pmatrix}\times2$
第3层	$\begin{pmatrix}3\times3,128\\3\times3,128\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\times3,128\\3\times3\times3,128\end{pmatrix}\times2$
第4层	$\begin{pmatrix}3\times3,256\\3\times3,256\end{pmatrix}\times2$	$\begin{pmatrix}3\times3\times3,256\\3\times3\times3,256\end{pmatrix}\times2$
第5层	$\begin{pmatrix}3\times3,512\\3\times3,512\end{pmatrix}\times2$ SE_block2，平均池化	$\begin{pmatrix}3\times3\times3,512\\3\times3\times3,512\end{pmatrix}\times2$ 平均池化

2.3　时间注意力约束

CEUS反映的病灶区信息主要与造影剂流入病灶区这一动态过程有关。对此，本文结合先验知识和数据时序特点，基于离散序列一阶差分和平滑L1范数设计了时间注意力约束损失函数，以约束模型对时间维度信息的学习过程。通过改变3D卷积核中时间维度参数，引导模型关注造影剂流入病灶区的时间窗口。

2.3.1　时序知识分析

CEUS中包含的视频亮度强化程度、亮度峰值等时间信息和造影剂流入病灶区阶段密切相关。在该阶段中，前后帧画面的变化反映了病灶区血管分布等空间信息，重点关注该时间段信息有利于模型高效地提取CEUS特征。视频画面亮度可以量化病灶区中造影剂含量，当造影剂流入病灶区时，画面亮度增加；造影剂流出病灶区时，亮度降低。抽取的32帧CEUS图像的亮度如图3（a）所示。模型应重点关注造影剂流入病灶区阶段，即绿色示意的帧数段。而蓝色帧数段对应的画面亮度逐渐降低，其变换趋势和模型所需关注的时间段内亮度变换趋势相反，一定程度上干扰模型的学习过程。此外，绿色帧数段含有的信息已经足够反映CEUS所包含的时空信息，而蓝色帧数段中的信息略显冗余。对此本文提出了时间注意力损失函数，以约束3D卷积核时间维度参数的优化过程，从而引导模型关注造影剂流入病灶区时间段的信息。

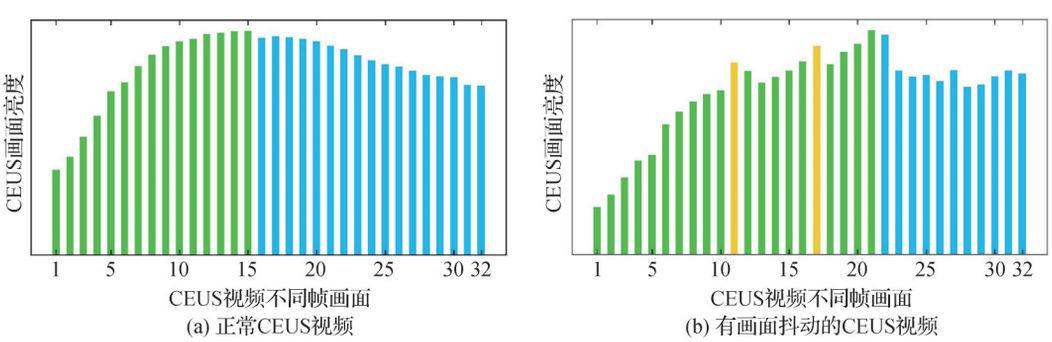


图3　时间注意力约束示意

Fig. 3　Illustration of temporal attention constraints ((a) CEUS normal video; (b) CEUS video with jitter)

2.3.2　时间注意力损失函数设计

本文先根据画面亮度序列的一阶差分值确定造影剂流入病灶区和流出病灶区的时间分界点，然后构造一个时间掩码。在此基础上，借助平滑L1范数设计时间注意力损失函数以约束模型的学习过程。

根据上述时序分析，在时间注意力分界点之前，画面亮度呈现递增趋势，之后亮度呈递减趋势，视频画面亮度这一离散序列的变换趋势可以用一阶差分表示。然而，由于CEUS视频画面抖动问题，尽管亮度总体呈现先升后降的趋势，但局部仍存在亮度波

动。如图3(b)所示,第11帧和第17帧处出现画面亮度波动。因此,仅根据单帧图像的一阶前向和后向差分值无法确定时间分界点的位置。对此,本文提出了如下方法判断时间注意力分界点,具体为

$$\boldsymbol{\Delta lm}_j = \boldsymbol{lm}_{j+1} - \boldsymbol{lm}_j \tag{1}$$

$$\boldsymbol{S}_j = \begin{cases} 1 & \boldsymbol{\Delta lm}_j > 0 \\ 0 & \boldsymbol{\Delta lm}_j \leqslant 0 \end{cases} \tag{2}$$

$$\boldsymbol{ATTN}_i = \begin{cases} 1 & \sum_{j=i}^{i+4} \boldsymbol{S}_j \geqslant 3 \\ 0 & \sum_{j=i}^{i+4} \boldsymbol{S}_j < 3 \end{cases} \tag{3}$$

$$TBP = \min\{i : \boldsymbol{ATTN}_i = 0\} \tag{4}$$

式中,$\boldsymbol{lm}$表示画面亮度序列,$\boldsymbol{lm}_j$表示第j帧CEUS图像的亮度。$\boldsymbol{\Delta lm}$为亮度差分序列,$\boldsymbol{\Delta lm}_j$表示第j帧CEUS图像的亮度一阶差分值。$\boldsymbol{S}$为亮度变换状态向量。TBP表示时间注意力分界点。

上述方法可根据一组CEUS视频图像亮度值确定一个时间注意力分界点,此分界点之前的帧图像是造影剂流入病灶区时间窗口内的画面,含有丰富的时空信息。由时序知识分析可知,模型需要忽略该点之后帧数图像,以消除噪声干扰。对此,本文构造了一个时间注意力约束损失函数,具体为

$$\boldsymbol{T}_{1_i}^{\mathrm{ATTN}} = \begin{cases} 0 & 1 \leqslant i < TBP \\ 1 & TBP \leqslant i \leqslant 32 \end{cases} \tag{5}$$

$$\boldsymbol{T}_{n_i}^{\mathrm{ATTN}} = \frac{1}{2^{n-2}} \sum_{j=1+2^{n-2}\times(i-1)}^{2^{n-2}\times i} \boldsymbol{T}_{1_j}^{\mathrm{ATTN}}, \quad n = 2,3,4,5 \tag{6}$$

$$\boldsymbol{NFM}_n = \boldsymbol{T}_n^{\mathrm{ATTN}} \odot \boldsymbol{FM}_n^{\mathrm{CEUS}}, \quad n = 1,2,3,4,5 \tag{7}$$

$$L_{\mathrm{temporal}} = \sum_{n=1}^{5} (SmoothL1(\boldsymbol{NFM}_n, \boldsymbol{O})) \tag{8}$$

式中,$\boldsymbol{T}_n^{\mathrm{ATTN}}$为第$n$层特征对应的时间掩码序列,$\boldsymbol{T}_{n_i}^{\mathrm{ATTN}}$为该掩码序列中第$i$个值,即第$n$层特征$T$维度上第$i$个特征图的掩码。$\boldsymbol{FM}_n^{\mathrm{CEUS}}$为CEUS分支提取到的第$n$层特征图。$\boldsymbol{NFM}_n$为第$n$层特征图中对应造影剂流出病灶区时间段的信息,即噪声信息。$SmoothL1$为平滑L1损失函数。时间注意力损失记为L_{temporal},通过引导$\boldsymbol{NFM}_n$置零的方式使得模型忽略造影剂流出病灶区时间段的信息,从而专注于造影剂流入病灶区时间段特征的提取。该损失函数可以对R(2+1)D网络中的时间维度卷积核的参数更新起约束作用,使得模型在时间维度上关注CEUS中有效信息。

2.4 特征融合模块

综合考虑B-US空间信息和CEUS时空信息有助于提升诊断的正确率,因此模型对双模态时空特征进行融合处理。为了实现维度不对等特征的融合处理,融合过程共分为两部分,先对双模态特征进行空间维度融合,然后在此基础上借助CEUS分支R(2+1)D三维网络完成时间维度上的特征融合。对于双模态特征空间维度的处理,通过双分支网络间横向连接的方式,使用特征融合模块实现B-US和CEUS特征空间维度上的融合。如图4所示,特征融合模块由空间特征融合模块和恒等映射分支组成。接下来,将详细阐述空间特征融合模块和恒等映射分支。

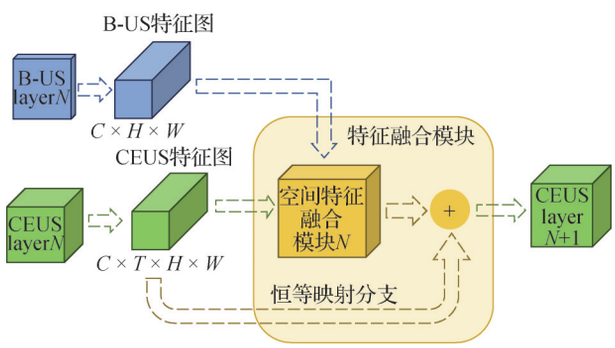


图4 特征融合模块

Fig. 4 Feature fuse block

2.4.1 空间特征融合模块

空间特征融合模块的结构如图5所示。首先,对CEUS特征进行维度调换,将$C \times T \times H \times W$的特征图转换为$T \times C \times H \times W$。随后,$T$维度上不同的CEUS特征图($C \times H \times W$)在通道维度上分别与B-US特征图($C \times H \times W$)进行拼接,得到$T \times 2C \times H \times W$的特征图。接下来,此特征图经过卷积处理,在融合双模态空间特征的同时减少特征通道数。考虑到CEUS中不同时刻对应的二维CEUS画面存在明显差异,T维度上不同的特征图($2C \times H \times W$)分别经过不同的卷积核进行处理,卷积核大小为$2C \times 1 \times 1$,处理后的特征图为$T \times C \times H \times W$。最后进行维度反调换,使空间特征融合模块输出的特征图维度为$C \times T \times H \times W$,与输入的CEUS特征图维度相等。

2.4.2 恒等映射分支

恒等映射分支的引入避免了特征融合过程中CEUS原本特征的损失。在特征融合模块中,第$N-1$层的CEUS特征图通过该分支与空间融合特征图进行对应位置加和处理,得到的新的特征图作为CEUS分支第N层特征提取网络的输入。恒等映射

分支使得第N层CEUS模型可以在第$N-1$层CEUS特征和融合特征中择优学习。同时恒等映射分支还保证输入到CEUS分支第N层的特征图有一部分直接来源于CEUS分支第$N-1$层提取到的特征。可以使得模型对CEUS的特征提取具有较好的连续性，从而更好地学习CEUS数据中包含的信息。

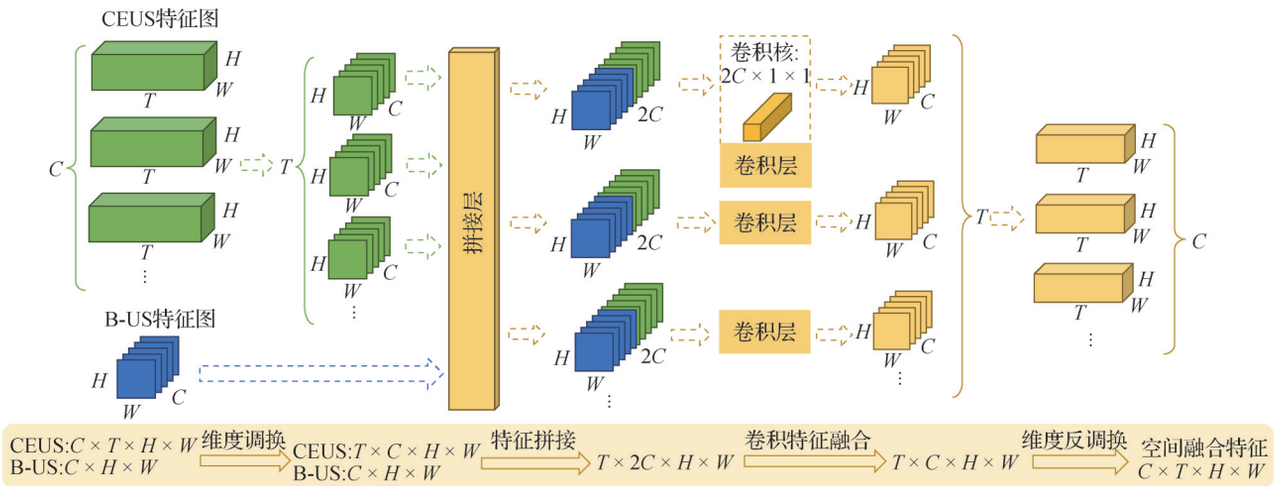


图5　空间特征融合模块

Fig. 5　Spatial feature fusion module

3　实　验

3.1　实验数据的获取和预处理

目前尚无同时包含乳腺肿瘤B-US和CEUS的公共数据集。本实验数据来源于中国医科大学附属盛京医院，共332个超声造影视频数据，分属于良性肿瘤、恶性肿瘤和炎症3个类别，其中，良性肿瘤101个，恶性肿瘤102个，炎症129个。造影视频画面的左半部分为CEUS，右半部分为B-US。

每个样本对应抽取32帧视频图像。原始视频帧图像经过裁剪，去除病人隐私信息，同时画面一分为二，一部分为CEUS，另一部分为B-US。接着CEUS图像被转换为灰度图像，在保证原有信息的同时减少CEUS特征提取分支第1层模型的参数量。然后对图像进行直方图均衡化和中值滤波处理，并通过等比放缩和黑色填充的方式统一尺寸为224×224像素。至此，共得到32对B-US和CEUS图像。

STFTCM的输入为1帧B-US图像和32帧CEUS图像。如图6所示，画面抖动问题使得同一个B-US视频不同帧画面间存在差异。本文视B-US画面抖动为一种数据增强的方式，对从1个视频中提取的32帧B-US图像进行子采样，得到2幅B-US图像。然后，这2幅B-US图像分别与32帧CEUS图像配对，从而将数据集扩充为原来的2倍。

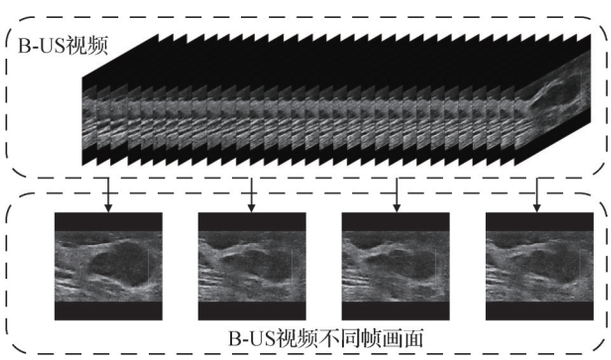


图6　B-US视频不同帧画面

Fig. 6　Different frames of B-US videos

3.2　实验设置

模型在PyTorch框架下实现，在NVIDIA GeForce RTX 3090 GPU上进行训练。批次大小为8，训练周期为100。模型训练过程选用动量随机梯度下降优化方法，动量值为0.9，权重衰减为0.005，学习率初始值设置为0.000 1，采用余弦退火衰减策略。

本文采用正确率、敏感性、特异性、宏平均F1值和曲线下面积(area under the curve，AUC)来评判模型的性能。数据共分3类，根据临床经验，在计算敏感性和特异性时，恶性肿瘤作为正类，其余两类为负类。宏平均F1值为3类情况各自F1的均值，在计算某一类F1时，将该类别当做正类，其余两类当做负类。

3.3 对比实验

本文共设置了3组对比实验，第1组使用B-US和CEUS双模态数据，另外两组分别使用B-US数据和CEUS数据。在3组对比实验中，首先对相应模态的原始数据进行视频帧采样处理，得到32帧视频画面图像。然后进行直方图均衡化和中值滤波处理，并统一图像尺寸大小为224×224像素。随后在模型训练过程中，为了缓解模型过拟合问题，将预训练的模型权重值作为初始值，并在预处理的数据集上迭代训练100次。最终得到对比实验结果如表2所示。3组对比实验结果表明了STFTCM性能优越，同时也验证了B-US和CEUS分支的出色性能。

表2 对比实验结果

Table 2 Comparative experimental results

	B-US	CEUS	模型	正确率	敏感性	特异性	宏平均F1	AUC
第1组（使用B-US和CEUS）	√	√	C3D(Tran等，2015)	0.853	0.636	**0.957**	0.845	0.929
	√	√	R(2+1)D(Tran等，2018)	0.794	0.818	0.826	0.796	0.864
	√	√	TRN(Zhou等，2018)	0.765	0.364	**0.957**	0.736	0.847
	√	√	TimeSformer(Bertasius等，2021)	0.529	0.364	0.783	0.504	0.818
	√	√	UniFormer(Li等，2022)	0.853	0.818	0.913	0.850	0.905
	√	√	TAdaFormer-B/16(Huang等，2023)	0.824	0.727	0.870	0.820	0.933
	√	√	TAdaFormer-L/14(Huang等，2023)	0.853	0.727	0.913	0.847	0.944
	√	√	本文	**0.882**	**0.909**	0.870	**0.883**	**0.952**
第2组（仅使用B-US）	√	–	C3D(Tran等，2015)	0.712	0.636	0.791	0.714	0.818
	√	–	R(2+1)D(Tran等，2018)	0.741	**0.673**	0.844	0.739	0.825
	√	–	TRN(Zhou等，2018)	0.706	0.363	**0.870**	0.679	0.818
	√	–	TimeSformer(Bertasius等，2021)	0.647	0.363	0.783	0.714	0.789
	√	–	UniFormer(Li等，2022)	0.735	0.546	**0.870**	**0.828**	0.692
	√	–	TAdaFormer-B/16(Huang等，2023)	0.735	0.636	0.783	0.732	0.821
	√	–	TAdaFormer-L/14(Huang等，2023)	0.706	0.545	0.782	0.703	0.815
	√	–	本文	**0.765**	0.636	0.826	0.758	**0.840**
第3组（仅使用CEUS）	–	√	C3D(Tran等，2015)	0.735	0.564	0.852	0.727	0.824
	–	√	R(2+1)D(Tran等，2018)	0.765	0.655	0.861	0.758	0.833
	–	√	TRN(Zhou等，2018)	0.747	0.636	0.861	0.761	0.828
	–	√	TimeSformer(Bertasius等，2021)	0.677	0.546	0.783	0.667	0.806
	–	√	UniFormer(Li等，2022)	0.759	0.693	0.852	0.832	0.773
	–	√	TAdaFormer-B/16(Huang等，2023)	0.794	0.727	0.826	0.800	0.837
	–	√	TAdaFormer-L/14(Huang等，2023)	0.794	**0.818**	0.783	0.791	0.832
	–	√	本文	**0.853**	0.727	**0.913**	**0.842**	**0.947**

注：加粗字体表示各组各指标最优结果，“√”表示使用对应数据，“–”表示不使用对应数据。

对于B-US和CEUS双模态数据，STFTCM在正确率、敏感性、宏平均F1和AUC指标上表现均好于其他先进模型。在模型预测正确率方面，STFTCM为88.2%，高于其他模型至少2.9%。同时STFTCM预测结果的敏感性为90.9%，远超其他模型。本文分析有以下原因：1）STFTCM整体结构设置更合理；2）模型进行了双模态特征融合处理；3）时间注意力损失约束模型关注有效信息。

在B-US分支的对比实验中，STFTCM的B-US分支模型性能同样表现优秀。B-US视频数据虽然有着丰富的空间信息，但是几乎不包含时间信息。用二维卷积网络针对视频中一幅图像提取特征可以减

少模型参数量，缓解因数据集过小引起的模型过拟合问题。

从CEUS分支的实验结果可得，STFTCM的CEUS特征提取分支在肿瘤预测方面同样优于其他模型。结果表明了时间注意力约束可以引导CEUS分支模型关注CEUS视频中特定时间段的信息。此外，实验结果还表明R(2 + 1)D网络由于其学习能力强且便于优化等优点，在数据集较小的情况下，训练后的模型性能较好。这也侧面验证了本文对CEUS分支主干网络选取的合理性。

3.4 消融实验

3.4.1 时间注意力损失函数消融实验

为验证时间注意力约束的有效性，本文对STFTCM和CEUS分支分别进行了时间注意力损失函数消融实验，并且根据CAM(class activation maps)图进行了相关的分析。

消融实验结果如表3所示，时间损失函数使得模型预测正确率升高5.8%，同时模型在宏平均F1、AUC、敏感性指标上均有所提升。原因有如下两点：1)时间注意力损失函数约束模型学习造影剂流入病灶区的时间窗口含有的时间信息；2)时间注意力损失函数引导模型关注造影剂流入病灶区阶段的帧间差异，而帧间亮度变换蕴含着病灶区血管分布等空间信息，所以时间注意力损失约束可以引导模型更好地获取空间维度信息。

表3 时间注意力损失函数消融实验结果

Table 3 Experimental results of temporal attention loss function ablation

B-US	CEUS	时间注意力约束	正确率	敏感性	特异性	宏平均F1	AUC
√	√	√	**0.882**	**0.909**	0.870	**0.883**	**0.952**
√	√	–	0.824	0.636	**0.913**	0.809	0.917
–	√	√	**0.853**	**0.727**	0.913	**0.842**	**0.947**
–	√	–	0.824	0.545	**0.957**	0.810	0.897

注：加粗字体表示对应数据下最优结果，“√”表示使用对应数据或约束，“–”表示不使用对应数据或约束。

为了验证上述观点，本文在有时间注意力损失函数约束和无时间注意力损失函数约束条件下分别进行模型训练，并在恶性肿瘤、良性肿瘤和炎症中分别选取1个样本，根据上述训练所得两种模型针对选取的3个样本，绘制了CAM图以及模型注意力随时间变化示意图，如图7所示。

从图7(a)可以看出，时间注意力约束会改变模型在空间维度上的注意力。该注意力约束引导模型关注造影剂流入病灶区时间段内前后帧之间的差异，并根据画面中的亮度变换获取病灶区血管分布等空间信息。而无时间约束的模型往往会忽略前后帧之间的变换，只是根据单帧画面的空间信息提取特征。

图7(b)验证了时间注意力损失函数会使模型在时间维度上重点关注视频亮度上升时间段的信息，从而有效地提取造影剂流入病灶区阶段的时间特征。而无时间注意力约束的模型对于视频的关注程度会随着视频亮度增加而增大，相比于有时间约束的模型，它更关注高亮画面的静态信息，而忽略了亮度变换这一过程中的动态信息。

由上述分析可推断出时间注意力约束有利于提升模型的性能。

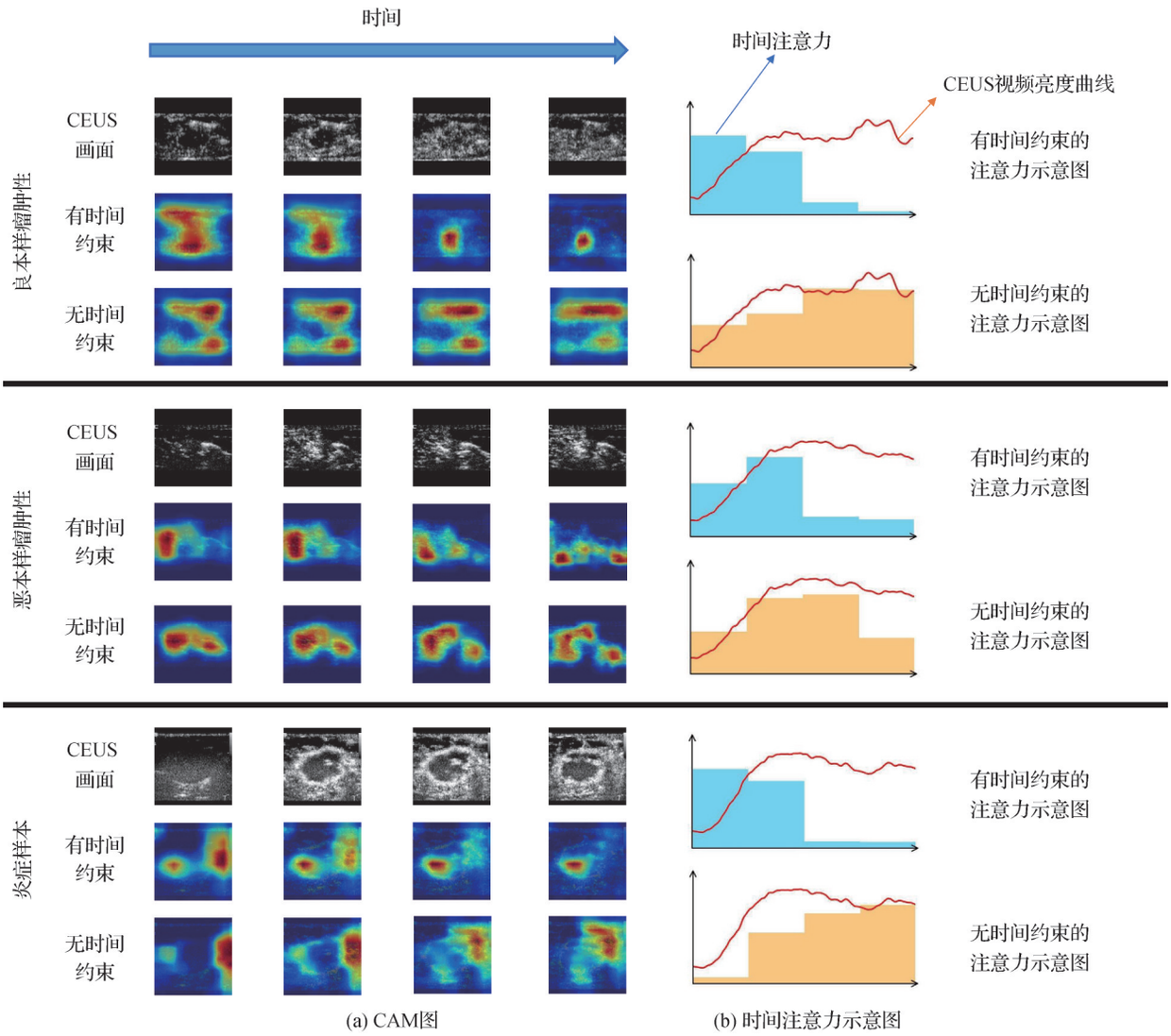
3.4.2 特征融合模块消融实验

为了验证特征融合模块的有效性，本文进行了特征融合模块消融实验，并绘制出ROC(receiver operating characteristic)曲线进行结果分析。

特征融合模块消融实验结果如表4所示。STFTCM双分支结构模型预测的正确率高于B-US单分支网络11.7%，高于CEUS分支2.9%。同时，双分支网络在敏感性、宏平均F1和AUC指标上的表现明显好于单分支网络。

如图8所示，相比于B-US单分支网络，双分支网络ROC曲线更靠近左上角，说明双分支网络预测正确率高于B-US单分支。相比之下，虽然CEUS分支的ROC曲线和双分支网络的ROC曲线更为相似，但双分支网络的AUC高于CEUS单分支模型，说明双分支网络的预测效果更好。

由此得出结论，特征融合模块可以有效地进行



(a) CAM图　　(b) 时间注意力示意图

图7　时间注意力损失函数消融实验所得CAM图和时间注意力示意图

Fig. 7　The CAM images and temporal attention diagram of the ablation experiment on temporal attention loss function ((a) CAM images; (b) temporal attention diagram of model)

表4　特征融合模块消融实验结果

Table 4　The results of feature fusion ablation experiments

B-US	CEUS	正确率	敏感性	特异性	宏平均F1	AUC
√	–	0.765	0.636	0.826	0.758	0.840
–	√	0.853	0.727	**0.913**	0.842	0.947
√	√	**0.882**	**0.909**	0.870	**0.883**	**0.952**

注：加粗字体表示各列最优结果。“√”表示使用对应数据，“–”表示不使用对应数据。

双模态信息的融合，有利于提高模型诊断的准确性。

3.5　特征融合模块实验

本文对特征融合模块进行了进一步的实验。实验结果证实了特征融合模块中的恒等映射分支有利于模型对CEUS特征的提取。考虑到不同时期的特征融合会影响到模型最终的诊断结果，本文针对特征融合时机也进行了相关的实验。

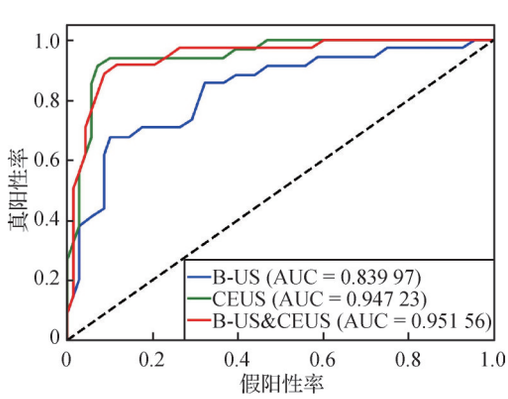


图8　B-US分支、CEUS分支和双分支网络的ROC曲线

Fig. 8　The ROC curves for the B-US branch, the CEUS branch, and the dual-branch network

首先，舍去特征融合模块中的恒等映射分支，使得CEUS分支的第N层模型的输入特征图完全来自于$N-1$层的特征融合模块。结果如表5所示，其预测正确率相较于采用恒等映射分支的模型降低了2.9%。分析原因如下，恒等映射分支可以使得CEUS分支第N层模型的输入特征图一部分直接来源于第$N-1$层CEUS分支的输出，保证了CEUS特征提取的连续性。同时恒等映射分支的引入可以使第N层的CEUS分支在第$N-1$层的融合特征和CEUS特征之间择优学习。

然后，为了探明STFTCM中最佳的特征融合时机，本文在双流网络不同的特征提取阶段进行特征融合。如表5所示，相比于早期特征融合，晚期的特征融合对于模型性能的提升更为显著。而同时使用早期融合和晚期融合对模型的性能提升最大。因为恒等映射分支的引入使得模型在双分支网络对应的4个层级之间自适应地进行特征融合，有利于提升诊断结果的准确性。

表5　特征融合模块恒等映射分支及特征融合时期实验结果

Table 5　The experimental results of feature fusion module identity mapping branch and feature fusion period

B-US & CEUS									
第1层	第2层	第3层	第4层	恒等映射分支	正确率	敏感性	特异性	宏平均F1	AUC
√	√	√	√	–	0.853	0.818	0.870	0.851	0.915
√	√	–	–	√	0.824	0.727	0.870	0.820	0.939
–	–	√	√	√	0.853	0.727	**0.913**	0.847	0.902
√	√	√	√	√	**0.882**	**0.909**	0.870	**0.883**	**0.952**

注：加粗字体表示各列最优结果，“√”表示含有特征融合模块或恒等映射分支，“–”表示不含有特征融合模块或恒等映射分支。

3.6　分类损失函数实验

为了探究训练过程中B-US分支分类损失函数和CEUS分支分类损失函数的引入是否能提升模型性能，本文对分类损失函数进行了相关的实验。实验中分别采用不同的分类损失函数组合进行模型的训练。

实验结果如表6所示，训练过程只用融合特征的分类损失所得模型的正确率为82.4%，而在此基础上添加CEUS分支分类损失可将正确率提升至85.3%。在融合特征分类损失的基础上加入CEUS分支分类损失可引导模型更好地提取CEUS特征，同时有助于双模态信息的融合。但是B-US分支分类损失和融合特征分类损失的组合使得模型预测正确率低于只用融合特征分类损失训练的模型。本文认为只引入B-US分支分类损失会使得模型更加关注于B-US特征的提取，而忽略了B-US特征和CEUS特征的融合过程。当同时使用3个分类损失函数时，模型既能聚焦于B-US和CEUS特征的获取，又能兼顾双模态特征之间的融合。因此训练得到的模型性能最佳，模型预测的正确率高达88.2%。

上述实验证明，模型训练阶段同时考虑CEUS分支分类损失、B-US分支分类损失和融合特征分类损失有助于引导模型更好地提取双模态信息，从而有利于后续的特征融合。

表6　分类损失函数实验结果

Table 6　The experimental results of the classification loss function

B-US & CEUS							
CEUS分类损失	B-US分类损失	融合特征分类损失	正确率	敏感性	特异性	宏平均F1	AUC
–	–	√	0.824	0.545	**0.957**	0.811	0.906
–	√	√	0.794	0.636	0.870	0.790	0.888
√	–	√	0.853	0.727	0.913	0.847	0.946
√	√	√	**0.882**	**0.909**	0.870	**0.883**	**0.952**

注：加粗字体表示各列最优结果，“√”表示使用对应损失，“–”表示不使用对应损失。

4 结 论

本文设计了一个融合时空特征与时间约束的双模态乳腺肿瘤诊断模型。STFTCM根据B-US和CEUS的信息维度特点,采用了异构双分支网络,在减少模型参数的同时,有效地从单帧B-US图像和一组CEUS视频画面提取到双模态信息。通过引入时间注意力损失函数引导模型关注CEUS中造影剂流入病灶区的时间窗口,充分提取时空信息。并且基于放射科医生双模态协同诊断的经验,设计了一个特征融合模块,将B-US中含有的二维空间信息通过横向连接的方式汇入CEUS分支,作为CEUS分支的补充信息。本文通过对比实验验证了STFTCM模型的优越性,并且通过消融实验等一系列实验证明了特征融合模块和时间注意力损失约束能有效地提升模型性能。

未来将在理解相关先验知识的基础上,继续挖掘数据中的信息,以提升模型性能。同时,也会继续收集规模更大且信息更多样的数据集,以满足模型训练的需求。

参考文献(References)

Anderson B O, Braun S, Lim S, Smith R A, Taplin S and Thomas D B. 2003. Early detection of breast cancer in countries with limited resources. The Breast Journal, 9(S2): S51-S59 [DOI: 10.1046/j.1524-4741.9.s2.4.x]

Balleyguier C, Opolon P, Mathieu M C, Athanasiou A, Garbay J R, Delaloge S and Dromain C. 2009. New potential and applications of contrast-enhanced ultrasound of the breast: own investigations and review of the literature. European Journal of Radiology, 69(1): 14-23 [DOI: 10.1016/j.ejrad.2008.07.037]

Belhumeur P N, Hespanha J P and Kriegman D J. 1997. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7): 711-720 [DOI: 10.1109/34.598228]

Bertasius G, Wang H and Torresani L. 2021. Is space-time attention all you need for video understanding? [EB/OL]. [2024-04-12]. https://arxiv.org/pdf/2102.05095.pdf

Byra M. 2018. Discriminant analysis of neural style representations for breast lesion classification in ultrasound. Biocybernetics and Biomedical Engineering, 38(3): 684-690 [DOI: 10.1016/j.bbe.2018.05.003]

Chen C, Wang Y, Niu J W, Liu X F, Li Q F and Gong X T. 2021. Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. IEEE Transactions on Medical Imaging, 40(9): 2439-2451 [DOI: 10.1109/TMI.2021.3078370]

Du J, Wang L, Wan C F, Hua J, Fang H, Chen J and Li F H. 2012. Differentiating benign from malignant solid breast lesions: combined utility of conventional ultrasound and contrast-enhanced ultrasound in comparison with magnetic resonance imaging. European Journal of Radiology, 81(12): 3890-3899 [DOI: 10.1016/j.ejrad.2012.09.004]

Evans A, Trimboli R M, Athanasiou A, Balleyguier C, Baltzer P A, Bick U, Camps Herrero J, Clauser P, Colin C, Cornford E, Fallenberg E M, Fuchsjaeger M H, Gilbert F J, Helbich T H, Kinkel K, Heywang-Köbrunner S H, Kuhl C K, Mann R M, Martincich L, Panizza P, Pediconi F, Pijnappel R M, Pinker K, Zackrisson S, Forrai G and Sardanelli F. 2018. Breast ultrasound: recommendations for information to women and referring physicians by the European Society of Breast Imaging. Insights into Imaging, 9(4): 449-461 [DOI: 10.1007/s13244-018-0636-z]

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

Gong R L, Shi J, Zhou W J and Wang C. 2022. Two-stage deep transfer learning for human breast ultrasound computer-aided diagnosis. Journal of Image and Graphics, 27(3): 898-910 (贡荣麟, 施俊, 周玮珺, 汪程. 2022. 面向乳腺超声计算机辅助诊断的两阶段深度迁移学习. 中国图象图形学报, 27(3): 898-910) [DOI: 10.11834/jig.210674]

Gong X, Yuan S, Xiang Y, Fan L and Zhou H. 2023. Domain knowledge-guided adversarial adaptive fusion of hybrid breast ultrasound data. Computers in Biology and Medicine, 164: #107256 [DOI: 10.1016/j.compbiomed.2023.107256]

Guo D H, Lu C Y, Chen D L, Yuan J Z, Duan Q M, Xue Z, Liu S X and Huang Y. 2024. A multimodal breast cancer diagnosis method based on knowledge-augmented deep learning. Biomedical Signal Processing and Control, 90: #105843 [DOI: 10.1016/j.bspc.2023.105843]

Guo R R, Lu G L, Qin B J and Fei B W. 2018. Ultrasound imaging technologies for breast cancer detection and management: a review. Ultrasound in Medicine and Biology, 44(1): 37-70 [DOI: 10.1016/j.ultrasmedbio.2017.09.012]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hooley R J, Scoutt L M and Philpotts L E. 2013. Breast ultrasonography: state of the art. Radiology, 268(3): 642-659 [DOI: 10.1148/radiol.13121606]

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and

Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Huang Z Y, Zhang S W, Pan L, Qing Z W, Zhang Y Y, Liu Z W and Ang M H Jr. 2023. Temporally-adaptive models for efficient video understanding [EB/OL]. [2024-04-12]. https://arxiv.org/pdf/2308.05787.pdf

Ilesanmi A E, Chaumrattanakul U and Makhanov S S. 2021. Methods for the segmentation and classification of breast ultrasound images: a review. Journal of Ultrasound, 24(4): 367-382 [DOI: 10.1007/s40477-020-00557-5]

Lee Y W, Huang C S, Shih C C and Chang R F. 2021. Axillary lymph node metastasis status prediction of early-stage breast cancer using convolutional neural networks. Computers in Biology and Medicine, 130: #104206 [DOI: 10.1016/j.compbiomed.2020.104206]

Li K C, Wang Y L, Gao P, Song G L, Liu Y, Li H S and Qiao Y. 2022. UniFormer: unified transformer for efficient spatiotemporal representation learning [EB/OL]. [2024-04-12]. https://arxiv.org/pdf/2201.04676.pdf

Lu S Y, Wang S H and Zhang Y D. 2022. SAFNet: a deep spatial attention network with classifier fusion for breast cancer detection. Computers in Biology and Medicine, 148: #105812 [DOI: 10.1016/j.compbiomed.2022.105812]

Luo L Y, Wang X, Lin Y, Ma X Q, Tan A D, Chan R, Vardhanabhuti V, Chu W C, Cheng K T and Chen H. 2024. Deep learning in breast cancer imaging: a decade of progress and future directions. IEEE Reviews in Biomedical Engineering, [DOI: 10.1109/RBME.2024.3357877]

Masud R, Al-Rei M and Lokker C. 2019. Computer-aided detection for breast cancer screening in clinical settings: scoping review. JMIR Medical Informatics, 7(3): #e12660 [DOI: 10.2196/12660]

Mo Y H, Han C, Liu Y, Liu M, Shi Z W, Lin J T, Zhao B C, Huang C W, Qiu B J, Cui Y F, Wu L, Pan X P, Xu Z Y, Huang X M, Li Z H, Liu Z Y, Wang Y and Liang C H. 2023. HoVer-trans: anatomy-aware HoVer-transformer for ROI-free breast cancer diagnosis in ultrasound images. IEEE Transactions on Medical Imaging, 42(6): 1696-1706 [DOI: 10.1109/TMI.2023.3236011]

Moon W K, Lee Y W, Ke H H, Lee S H, Huang C S and Chang R F. 2020. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Computer Methods and Programs in Biomedicine, 190: #105361 [DOI: 10.1016/j.cmpb.2020.105361]

Qiu Z F, Yao T and Mei T. 2017. Learning spatio-temporal representation with pseudo-3D residual networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5534-5542 [DOI: 10.1109/ICCV.2017.590]

Siegel R L, Giaquinto A N and Jemal A. 2024. Cancer statistics, 2024. CA: A Cancer Journal for Clinicians, 74(1): 12-49 [DOI: 10.3322/CAAC.21820]

Singh S P, Wang L P, Gupta S, Goli H, Padmanabhan P and Gulyás B. 2020. 3D deep learning on medical images: a review. Sensors, 20(18): #5097 [DOI: 10.3390/s20185097]

Tran D, Bourdev L, Fergus R, Torresani L and Paluri M. 2015. Learning spatiotemporal features with 3D convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 4489-4497 [DOI: 10.1109/ICCV.2015.510]

Tran D, Wang H, Torresani L, Ray J, LeCun Y and Paluri M. 2018. A closer look at spatiotemporal convolutions for action recognition// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6450-6459 [DOI: 10.1109/CVPR.2018.00675]

Wubulihasimu M, Maimaitusun M, Xu X L, Liu X D and Luo B M. 2018. The added value of contrast-enhanced ultrasound to conventional ultrasound in differentiating benign and malignant solid breast lesions: a systematic review and meta-analysis. Clinical Radiology, 73(11): 936-943 [DOI: 10.1016/j.crad.2018.06.004]

Yang Z Q, Gong X, Zhu D and Guo Y. 2020. Cooperative suppression network for bimodal data in breast cancer classification. Journal of Image and Graphics, 25(10): 2218-2228 (杨子奇, 龚勋, 朱丹, 郭颖. 2020. 乳腺超声双模态数据的协同约束网络. 中国图象图形学报, 25(10): 2218-2228) [DOI: 10.11834/jig.200246]

Zhao X, Gong X, Fan L and Luo J. 2022. Attention-based networks of human breast bimodal ultrasound imaging classification. Journal of Image and Graphics, 27(3): 911-922 (赵绪, 龚勋, 樊琳, 罗俊. 2022. 结合注意力机制的乳腺双模态超声分类网络. 中国图象图形学报, 27(3): 911-922) [DOI: 10.11834/jig.210370]

Zhou B L, Andonian A, Oliva A and Torralba A. 2018. Temporal relational reasoning in videos//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 831-846 [DOI: 10.1007/978-3-030-01246-5_49]

Zhu Y, Li X Y, Liu C H, Zolfaghari M, Xiong Y J, Wu C R, Zhang Z, Tighe J, Manmatha R and Li M. 2020. A comprehensive study of deep video action recognition [EB/OL]. [2024-04-12]. https://arxiv.org/pdf/2012.06567.pdf

Zonderland H M. 2000. The role of ultrasound in the diagnosis of breast cancer. Seminars in Ultrasound, CT and MRI, 21(4): 317-324 [DOI: 10.1016/S0887-2171(00)90026-X]

作者简介

李一宸，男，本科生，主要研究方向为计算机视觉和医学图像处理。E-mail: ddzdhlyc2021@126.com

陈大力，通信作者，男，教授，主要研究方向为计算机视觉和深度学习。E-mail: chendali@ise.neu.edu.cn

郭丁豪，男，博士研究生，主要研究方向为计算机视觉和医学图像处理。E-mail: 640307541@qq.com

孙羽，男，本科生，主要研究方向为计算机视觉和医学图像处理。E-mail: 20215311@stu.neu.edu.cn