doi：10. 12068/j. issn. 1005－3026. 2026. 20250067

基于面部视频的非接触式血氧饱和度估计方法

齐 林[1,2,3]，高启赫[1]，关舒月[1]，李永春[4]

（1. 东北大学 医学与生物信息工程学院，辽宁 沈阳 110169；2. 东北大学 医学影像计算教育部重点实验室，辽宁 沈阳 110169；3. 东北大学 医学成像与智能分析教育部工程研究中心，辽宁 沈阳 110169；4. 沈阳康泰电子科技股份有限公司，辽宁 沈阳 110167）

摘 要：针对远程光电容积描记法（rPPG）在非接触式血氧饱和度（SpO_2）测量中存在的时空特征建模不足以及复杂场景下鲁棒性差的挑战，提出了一种趋势感知时空融合网络（trend-aware spatio-temporal fusion network, TAST-Net）. 该网络通过一个创新的双路融合架构，将3D卷积神经网络（3D CNN）分支提取的局部生理特征与ViViT（video vision transformer）分支捕捉的全局时空依赖进行协同融合. 为增强模型对信号动态变化的敏感性，设计了一种结合均方误差与皮尔逊相关性损失的加权组合损失函数. 在2个公开数据集上的实验结果表明，TAST-Net表现出优秀的性能：在PURE（pulse rate estimation）数据集上均方根误差（e_{RMS}）为0. 53%，平均绝对误差（e_{MA}）为0. 37%，皮尔逊相关系数（R）为0. 96；在更具挑战性的VIPL-HR（visual information processing and learning-heart rate）数据集上，e_{RMS}为0. 84%，e_{MA}为0. 57%，R为0. 82，其综合性能优于其他对比方法. 研究结果表明，TAST-Net为从面部视频中实现准确、稳健的SpO_2估计提供了一个有效的方案，并验证了融合局部与全局特征策略在rPPG信号处理中的有效性.

关 键 词：远程光电容积描记法；深度学习；非接触；血氧饱和度估计；面部视频

中图分类号：TP 391.41；R 318.08 **文献标志码：**A **文章编号：**1005-3026（2026）01-0042-10

Non-contact Estimation Method of Blood Oxygen Saturation Based on Facial Videos

QI Lin[1,2,3], *GAO Qi-he*[1], *GUAN Shu-yue*[1], *LI Yong-chun*[4]

（1. College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110169, China; 2. Key Laboratory of Medical Image Computing, Ministry of Education, Northeastern University, Shenyang 110169, China; 3. Engineering Research Center of Medical Imaging and Intelligent Analysis, Ministry of Education, Northeastern University, Shenyang 110169, China; 4. Shenyang Contain Electronic Technology Co., Ltd., Shenyang 110167, China. Corresponding author: LI Yong-chun, E-mail: liyongchun@contain.com.cn）

Abstract: To address the challenges of inadequate spatio-temporal feature modeling and poor robustness in complex scenarios for non-contact blood oxygen saturation（SpO_2）measurement using remote photoplethysmography（rPPG）, a trend-aware spatio-temporal fusion network（TAST-Net）was proposed. The proposed network adopted an innovative dual-branch fusion architecture that synergistically fused local physiological features extracted by a 3D convolutional neural network（3D CNN）branch with global spatio-temporal dependencies captured by a video vision transformer（ViViT）branch. To enhance the model's sensitivity to signal dynamics, a weighted composite loss function combining mean squared error（MSE）and Pearson correlation loss was designed. Experimental results on two public datasets demonstrate the superior performance of TAST-Net. On the pulse rate estimation（PURE）dataset, it achieves a root mean squared error（e_{RMS}）of 0. 53%, a mean absolute error（e_{MA}）of 0. 37%, and a Pearson correlation coefficient（R）of 0. 96. On the more challenging visual information processing and learning-heart rate（VIPL-HR）dataset, the e_{RMS}, e_{MA}, and R reach 0. 84%, 0. 57%, and 0. 82,

收稿日期：2025-06-12
基金项目：辽宁省重点研发项目（2024JH2/102500076）.
作者简介：齐 林（1981—），男，吉林长春人，东北大学副教授.
通信作者：李永春，E-mail：liyongchun@contain.com.cn.

respectively, outperforming other comparative methods. These findings indicate that TAST-Net provides an effective solution for accurate and robust SpO_2 estimation from facial videos and validates the advantage of integrating local and global features in rPPG signal processing.

Key words: remote photoplethysmography; deep learning; non-contact; blood oxygen saturation estimation; facial video

血氧饱和度(SpO_2)作为衡量血液中氧合血红蛋白比例的关键指标,反映了肺部气体交换和心脏循环功能的综合状态,是评估个体呼吸与循环系统健康状况的重要生理参数.在医学领域,SpO_2的异常往往与呼吸系统疾病、心血管疾病和睡眠呼吸暂停[1]等健康问题密切相关,其持续下降是疾病恶化的重要生理信号.近年来,随着COVID-19等呼吸道疾病的全球流行,人们对远程健康监测的需求不断增长,精准、便捷的SpO_2监测技术显得尤为重要[2].

传统的SpO_2测量方法主要包括指夹式脉搏血氧仪[3]和血气分析法.指夹式脉搏血氧仪利用光电容积描记法(PPG)[4]测量SpO_2,虽便携易用,但在低灌注、晃动时存在局限性,且接触式测量不适用于某些患者或场景.血气分析法虽为"金标准",但属有创检测,且无法连续监测.因此,研究非接触式的SpO_2测量方法至关重要.

随着计算机视觉和人工智能技术的发展,远程光电容积描记法(rPPG)为非接触式估计提供了新的可能性.该技术的发展大致遵循从传统信号处理到深度学习的演进路径.在早期阶段,研究主要集中于传统信号处理方法.自Verkruysse等[5]的开创性工作后,发展出了基于颜色空间(如颜色空间方法(CHROM)[6])、盲源分离(如独立成分分析(ICA)[7]、主成分分析(PCA)[8])及物理反射模型(如平面正交投影(POS)[9])等多种方法.这些方法为rPPG奠定了理论基础,但其性能高度依赖于理想环境假设和手动设计的特征,在面对真实场景中的运动和光照变化时,鲁棒性与泛化能力较弱.

为克服传统方法的局限性,端到端的深度学习模型逐渐成为研究主流,极大地推动了技术的发展.学者首先引入了卷积神经网络(CNN)[10-11],利用其强大的局部特征提取能力来自动学习rPPG信号的时空模式.随后,为解决CNN在捕捉长程依赖上的不足,学者进一步引入了Transformer架构[12-13],利用其自注意力机制来建模视频帧间的全局关联.此外,还衍生出了结合少量标签的弱监督学习[14]与不依赖标签的无监督对比学习[15]等范式,以降低对大规模标注数据的依赖.

尽管深度学习方法取得了显著进展,但现有方法仍面临两大核心挑战:时空特征建模不足与复杂场景鲁棒性差.一方面,在时空建模上,单一模型范式难以在全局依赖和局部细节之间取得平衡.例如,基于CNN的方法虽擅长提取局部特征,但其有限的感受野难以整合空间上离散的面部区域以构建全局生理图谱;而基于Transformer的方法虽能捕捉长程依赖,却可能在缺乏有效先验时忽略对rPPG至关重要的像素级精细颜色变化.另一方面,在鲁棒性上,面部视频中的rPPG信号极其微弱,在真实场景中极易被头部运动、光照变化等强噪声淹没.这些噪声与真实生理信号在频域上常常发生混叠,导致模型难以有效区分.因此,本文提出一种创新的双路融合架构,旨在协同2种架构的优势,以期在复杂场景下实现更高的估计精度与鲁棒性.

基于上述分析,本研究的主要贡献如下:首先,提出一种基于趋势感知时空融合网络的非接触式SpO_2估计模型.该模型利用3D卷积神经网络(3D CNN)和ViViT的优势互补,构建了端到端的双路融合架构:3D CNN分支负责提取并细化局部生理信号特征,而ViViT分支则捕捉视频的全局时空特征.通过双路特征的有效融合,实现了对生理信号细节与视频长程时空特征的协同作用,旨在提高SpO_2估计的精度与可靠性;其次,为进一步优化模型性能,研究中设计了一种加权组合损失函数,该函数结合了均方误差(MSE)损失与皮尔逊相关系数损失的特点,不仅关注估计数值的准确性,更致力于提升模型对生理信号动态变化模式的捕捉能力;最后,本文在PURE和VIPL-HR这2个公开数据集上设计并进行了一系列对比实验,将TAST-Net与多种深度学习模型进行比较,从而验证所提出的TAST-Net模型能够在较复杂的场景下实现准确、稳健的SpO_2估计.

1　面向非接触式血氧饱和度估计的TAST-Net模型

1.1　模型概述

目前基于rPPG技术估计人体血氧饱和度的

研究虽取得了一定进展,但仍存在一系列应用层面的局限性与挑战.面部视频中反映的生理信号非常微弱,易受头部运动、光照变化、肤色等因素干扰,导致信号质量差、信噪比低,从而在复杂的实际场景中难以实现对生理信号的准确估计.现有深度学习方法中,卷积神经网络(CNN)和Transformer[16]模型被广泛应用于图像和视频数据的特征提取,并执行分类、识别和预测等任务.然而,CNN通常用于提取局部空间特征,其感受野有限,难以捕捉时间维度上的全局依赖性.Transformer在序列建模方面表现出色,但如果缺乏有效的局部空间特征,易受到与血氧无关或误导性的时间特征影响.

针对这些挑战,本文提出一种基于3D CNN和ViViT[17]的双路融合的非接触式血氧饱和度估计的网络模型TAST-Net.TAST-Net通过3D CNN分支对视频逐步提取并细化局部生理信号特征,有效捕捉时序维度的生理信息;同时,ViViT分支利用Tubelet Embedding结构嵌入位置编码(Positional Encoding),捕捉视频的全局时空特征.双路提取的特征经融合后输入多层感知机(MLP)回归头,实现对生理信号细节与视频长程时空特征的协同作用,从而提高SpO_2估计的精度与可靠性.

1.2 TAST-Net模型结构

为实现局部生理细节与全局时空动态的协同建模,本文设计了TAST-Net模型,其整体网络框架如图1所示.该框架由1个负责提取局部时空特征的3D CNN分支和1个负责捕捉长程依赖关系的ViViT分支并行构成.2个分支的输出特征最终被融合,并通过1个多层感知机回归头来估计SpO_2值.

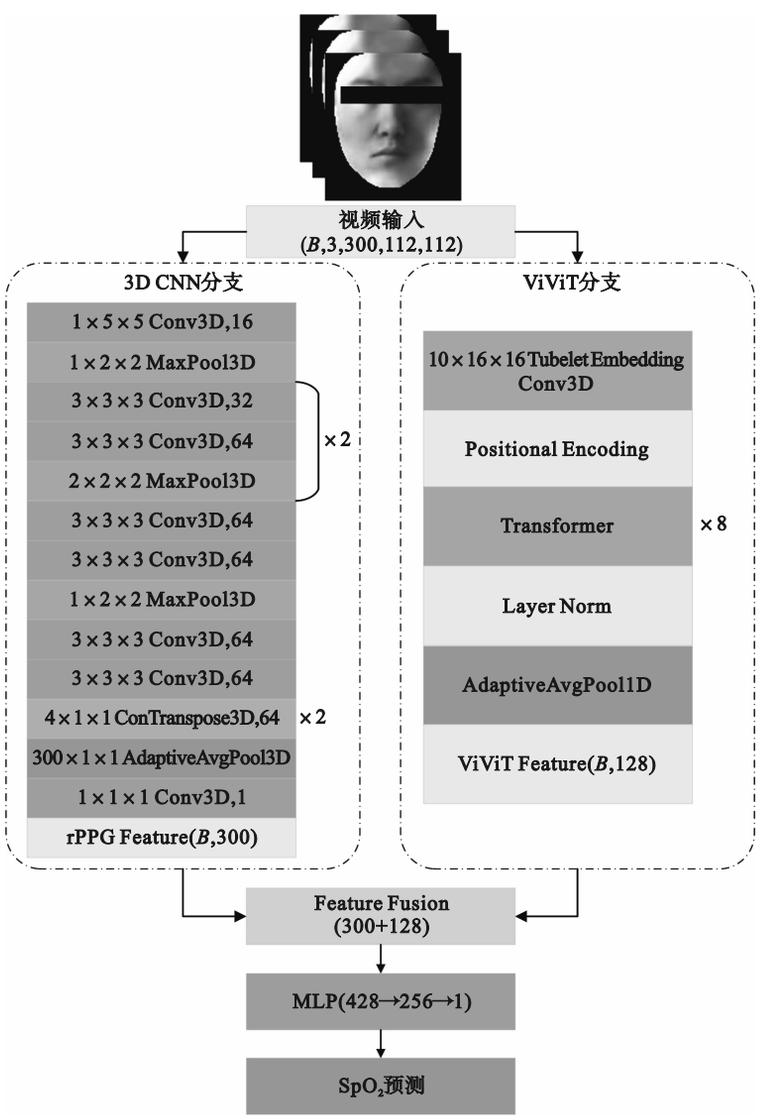


图1 TAST-Net网络框架

Fig. 1 Architecture of TAST-Net

1) 模型输入. TAST-Net模型的输入是维度为(B,3,300,112,112)的视频片段,其中B为批处理大小(batch size,本文设为8),3为RGB(red green blue)颜色三通道,300为时间维度(帧),112像素×112像素为每帧的空间分辨率.

2) 3D CNN分支:局部时空特征提取. 该分支的核心任务是从输入视频中提取与rPPG信号相关的局部时空特征.其结构设计通过堆叠的3D卷积层逐步扩大感受野以提取深层特征,并利用池化层进行下采样以降低计算复杂度.

该分支首先通过1个核尺寸为$1\times5\times5$的Conv3D层进行初步的空间特征提取,随后利用1个核尺寸为$1\times2\times2$的MaxPool3D层将特征图空间尺寸减半,维度从(B,3,300,112,112)变为(B,16,300,56,56).

接下来,数据流经过一系列3D卷积与最大池化层的交替堆叠以提取更高级的时空特征.首先进入1个重复堆叠2次的特征提取模块:该模块通过连续的3×3×3的Conv3D层并将通道数加深至64,并利用1个2×2×2的MaxPool3D层进行时空下采样.重复2次后,特征图维度降至(B,64,75,14,14).随后经2个连续的3×3×3的Conv3D层进一步提取特征,并利用1个1×2×2的MaxPool3D层对空间维度进行下采样.然后通过2个连续的3×3×3的Conv3D层进一步提取特征,获得维度为(B,64,75,7,7)的深层特征表示.

为恢复在时序下采样中损失的部分时间信息,模型采用2个核尺寸为$4\times1\times1$的转置卷积层(ConvTranspose3D),将时间维度上采样至300,输出维度调整至(B,64,300,7,7).随后,1个自适应平均池化层(AdaptiveAvgPool3D)在空间维度上进行全局池化,得到(B,64,300,1,1)的时序特征.最后,通过1个$1\times1\times1$的Conv3D层将通道数降维至1,并重塑(reshape)为维度(B,300)的rPPG特征序列.

3) ViViT分支:全局时空依赖建模. 该分支旨在利用Transformer架构[16]捕捉视频帧间的长程依赖关系和全局上下文信息,以弥补CNN局部感受野的不足.

该分支首先通过1个Tubelet Embedding模块将输入视频进行序列化.该模块使用(10,16,16)的3D卷积核将视频分割成1 470个“管状”片段(tubelets),并将每个tubelet线性映射为1个128维的嵌入向量.

这些嵌入向量在添加了可学习的位置编码以保留其原始时空信息后,被送入1个由8层标准Transformer编码器组成的模块.通过多头自注意力机制,模型能够对所有tubelets之间的依赖关系进行建模,有效捕捉视频的全局动态信息.

经过Transformer编码器处理后,输出特征首先经过层归一化(Layer Norm)处理,随后通过1个一维自适应平均池化层(AdaptiveAvgPool1D)对序列维度进行全局池化,将序列特征聚合成1个维度为(B,128)的固定长度特征向量,作为该分支的最终输出.

4) 特征融合与回归估计. 最终,3D CNN分支输出的局部时序特征(B,300)与ViViT分支输出的全局特征(B,128)在特征维度上进行拼接(concatenation),形成一个维度为(B,428)的融合特征向量.该向量随后被送入1个由2个全连接层组成的多层感知机回归头,将特征维度从428映射到256,再最终映射到1,实现对SpO_2值的端到端估计.

1.3　损失函数设计

为精确地从输入面部视频中估计血氧饱和度,训练的核心在于优化本文设计的组合损失(L_T)函数,该损失函数是均方误差损失(L_{MSE})和负皮尔逊相关系数损失($L_{1-Pearson}$)的加权组合.

均方误差损失定义为

$$L_{MSE}=\frac{1}{N}\sum_{i=1}^{N}\left(SpO_{2\,pre}^{(i)}-SpO_{2\,gt}^{(i)}\right)^2; \tag{1}$$

负皮尔逊相关系数损失定义为

$$L_{1-Pearson}=1-\frac{\sum_{i=1}^{N}\left(SpO_{2\,pre}^{(i)}-\overline{SpO_2}_{pre}\right)\left(SpO_{2\,gt}^{(i)}-\overline{SpO_2}_{gt}\right)}{\sqrt{\sum_{i=1}^{N}\left(SpO_{2\,pre}^{(i)}-\overline{SpO_2}_{pre}\right)^2}\sqrt{\sum_{i=1}^{N}\left(SpO_{2\,gt}^{(i)}-\overline{SpO_2}_{gt}\right)^2}}. \tag{2}$$

本文设计的组合损失定义为

$$L_T=\alpha L_{MSE}+\beta L_{1-Pearson}. \tag{3}$$

其中:$SpO_{2\,pre}^{(i)}$和$SpO_{2\,gt}^{(i)}$分别代表估计的SpO_2值和实际测量SpO_2的真值;$\overline{SpO_2}_{pre}$和$\overline{SpO_2}_{gt}$分别代表估计的SpO_2和实际测量SpO_2的均值;N代表样本总数;α和β代表平衡两部分损失的权重因子.

2　实验设计与结果分析

为了全面评估本文所提出的TAST-Net模型

的有效性,在PURE[18]和VIPL-HR[19]这2个公开数据集上进行了实验验证,并与多种基于rPPG的血氧饱和度估计算法进行了对比分析.本实验旨在探讨各方法在不同测试条件下的泛化能力、估计误差及相关性.

2.1 实验数据集

本文选用了2个在rPPG研究领域广泛使用的公开数据集:PURE和VIPL-HR.

PURE数据集[18]:该数据集主要用于评估运动伪影对rPPG信号的影响,包含10名受试者(8男2女)的面部视频数据.实验设置了6种不同的头部运动场景:静止、慢速平移、快速平移、慢速旋转、中速旋转和说话.视频以640像素×480像素的分辨率、30 fps帧率录制,并同步记录了心率(HR)、血氧饱和度(SpO_2)和血液容积脉搏(BVP)信号作为参考值.该数据集的特点是场景控制良好,但运动干扰较为明显,适合测试模型对头部运动的鲁棒性.

VIPL-HR数据集[19]:该数据集是1个规模更大、场景更复杂的多模态数据库,包含107名受试者(79男28女)的2 378个视频片段.该数据集旨在模拟更真实的、约束较少的应用场景,涵盖9种不同的场景,包括静止、头部运动(向上/向下、向左/向右、自由运动)、不同光照条件(暗光、强光)以及不同设备的采集方式.视频分辨率和帧率不一,同步记录了HR, SpO_2和BVP信号.VIPL-HR数据集的规模和复杂性为评估模型在不同光照、运动和设备条件下的泛化能力提供了良好的基础.

2.2 数据预处理

为了确保模型能够专注于面部区域的生理信号,并提升信号质量,本文对所有视频帧进行了统一的预处理.

1)人脸检测与感兴趣区域(ROI)提取:本研究使用dlib库[20]进行人脸检测与姿态对齐.基于检测到的68个面部关键点(见图2a),通过几何变换对人脸姿态进行校正,并提取1个尺寸标准化为112像素×112像素的ROI.该对齐过程确保了包括额头在内的完整面部结构被包含并位于ROI内(见图2b),为后续处理提供了统一输入.

2)背景去除:在获得对齐的ROI后,为彻底分离面部皮肤与头发、衣领等背景噪声,本文采用了Google的MediaPipe Face Mesh技术[21]进行第二阶段的精细化分割.该技术可生成1个包含468个关键点的密集面部网格,其轮廓紧密贴合从下巴至发际线的完整面部边界.利用该网格的外圈轮廓点生成一个精确的面部多边形掩码(mask),并将其应用于ROI,从而得到如图2c所示的纯净面部图像.

3)欧拉视频放大:由于rPPG信号非常微弱,本文采用EVM(Eulerian video magnification)算法[22]进行视频帧的颜色放大.该技术通过对视频进行拉普拉斯金字塔分解,并沿时间轴对特定频带内的信号进行放大,从而增强皮肤区域因血流变化引起的微弱颜色变化.本研究设置放大倍数为120,并选择0.4~4 Hz的频率范围以匹配心率波动.处理后的ROI如图2d所示.

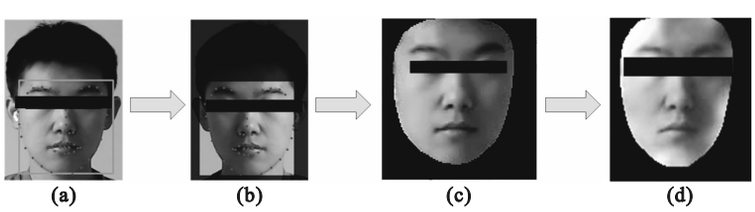


图2 数据预处理流程

Fig. 2 Data preprocessing process

(a)—面部关键点检测;(b)—提取的面部ROI;(c)—对ROI进行背景去除;(d)—欧拉视频放大后的面部.

2.3 实验配置

在所有实验中,输入视频片段的尺寸统一调整为3×300×112×112(视频通道数×帧数×帧高(像素)×帧宽(像素)).Batch size设置为8,训练阶段用60个epoch训练模型.学习率设置采用了动态调整策略,初始学习率(learning rate)设置为0.000 1.当验证损失连续5个epoch未下降(patience=5)时,学习率自动缩小为当前值的1/10.Transformer的层数设置为N=8.损失函数中的权重α设为0.1,β设为0.9.本研究使用Pytorch框架实现,并使用AdamW优化器[23]在NVIDIA RTX 4090 GPU上进行训练.在测试阶段,使用10 s(300帧)的视频片段来估计SpO_2值.

尽管预训练能够在某些情况下加速收敛并

提升性能,但其效果依赖于源数据集与目标任务之间的相关性.本研究重点在于验证TAST-Net架构本身的有效性,而非依赖额外的大规模预训练知识,本文选择对3D CNN和ViViT分支均采用从零开始(from scratch)的训练策略.在参数初始化方面,采用了深度学习领域常用的Kaiming[24]初始化方法,以保证网络在训练初期的稳定性和收敛性.

2.4　评估指标

本研究选取3个评估指标来比较SpO_2估计值和标签之间的误差:

1)均方根误差e_{RMS}(root mean squared error):

$$e_{RMS}=\sqrt{\left[\sum_{i=1}^{N}\left(SpO_{2\,pre}^{(i)}-SpO_{2\,gt}^{(i)}\right)^2\right]/N}\ ; \tag{4}$$

2)平均绝对误差e_{MA}(mean absolute error):

$$e_{MA}=\frac{1}{N}\sum_{i=1}^{N}\left|SpO_{2\,pre}^{(i)}-SpO_{2\,gt}^{(i)}\right|; \tag{5}$$

3)皮尔逊相关系数(R):

$$R=\frac{\sum_{i=1}^{N}\left(SpO_{2\,pre}^{(i)}-\overline{SpO_2}_{pre}\right)\left(SpO_{2\,gt}^{(i)}-\overline{SpO_2}_{gt}\right)}{\sqrt{\sum_{i=1}^{N}\left(SpO_{2\,pre}^{(i)}-\overline{SpO_2}_{pre}\right)^2}\sqrt{\sum_{i=1}^{N}\left(SpO_{2\,gt}^{(i)}-\overline{SpO_2}_{gt}\right)^2}}. \tag{6}$$

2.5　训练策略

在实验中,2个数据集均按照6:2:2的比例被随机划分为训练集、验证集和测试集.

模型训练共进行60个周期(epochs),批处理大小设为8.优化器采用AdamW优化器[23],并使用本文提出的组合损失(L_T)函数作为优化目标.为实现动态学习率调整,本文采用了ReduceLROnPlateau调度策略[25]:当验证集上的损失值连续5个周期未下降时,学习率将自动衰减为原先的1/10,以促进模型在训练后期进行更精细的参数搜索.

在每个周期训练结束后,模型在验证集上进行性能评估.依据验证集上取得的最大皮尔逊相关系数(R)作为唯一标准,来保存性能最佳的模型权重.整个训练过程完成后,该最佳模型将在独立的测试集上进行最终的性能评估,所采用的评估指标为e_{RMS},e_{MA}和R.

2.6　对比模型

为了全面评估TAST-Net模型的性能,本研究选择了以下4种深度学习模型进行对比:

1)3D-CNN[26]:一种作为对比基准的端到端3D卷积网络.该模型采用浅层的压缩式架构,通过卷积与池化层对视频进行连续时空下采样,旨在将信息最终聚合为单一特征向量用于直接回归.其设计与本研究TAST-Net中的3D CNN分支有显著不同:后者是一个更深层的特征提取网络,采用"降采样-上采样"的恢复式结构来精细化并保留完整的时序动态信息,而非进行纯粹的信息压缩.

2)MultiPhysNet[27]:一种专为多种生理信号(包括SpO_2)估计而设计的深度神经网络,其结构考虑了多任务学习的特点.

3)ITSCAN[28](innovative temporal shift coordinate attention network):一种基于时间位移模块的神经网络模型,其核心特点是包含一个提取时空特征的运动分支和一个利用坐标注意力机制处理特征通道与位置信息的外观分支,专为远程生理信号监测设计.

4)MMFM[29](multi-model fusion method):一种多模型融合方法,它结合了基于颜色通道重建信号的颜色通道模型(CCM)和基于深度学习网络的模型(NBM),旨在充分利用面部视频中的颜色和深层特征信息进行SpO_2估计.

2.7　实验结果与分析

本文将TAST-Net与4种深度学习模型(3D-CNN, MultiPhysNet, ITSCAN 和 MMFM)在PURE和VIPL-HR公开数据集上进行了全面评估.所有方法的性能指标情况见表1和表2.

表1　不同模型在PURE数据集上的血氧饱和度估计结果

Table 1　Blood oxygen saturation estimation results of different models on PURE dataset

模型	e_{RMS}/%	e_{MA}/%	R	参数量/10^6	推理速度/(ms·帧$^{-1}$)
3D-CNN	6.13	5.89	0.25	13.33	1.182
MultiPhysNet	0.91	0.72	0.86	<u>0.88</u>	0.153
ITSCAN	1.72	1.36	0.73	22.48	0.194
MMFM	<u>0.89</u>	<u>0.66</u>	<u>0.87</u>	**0.74**	**0.077**
Our TAST-Net	**0.53**	**0.37**	**0.96**	3.24	<u>0.149</u>

注:加粗数值表示最优结果,下划线数值表示次优结果.下同.

表2 不同模型在VIPL-HR数据集上的血氧饱和度估计结果

Table 2 Blood oxygen saturation estimation results of different models on VIPL-HR dataset

模型	e_{RMS}/%	e_{MA}/%	R	参数量/10^6	推理速度/(ms·帧$^{-1}$)
3D-CNN	2.62	2.42	0.63	13.33	1.182
MultiPhysNet	0.90	0.66	0.80	0.88	0.153
ITSCAN	1.14	0.72	0.70	22.48	0.194
MMFM	1.16	0.87	0.59	**0.74**	**0.077**
Our TAST-Net	**0.84**	**0.57**	**0.82**	3.24	0.149

由表1和表2的实验结果可知,本文提出的TAST-Net模型在各项评估指标上均展现了显著的优势.

PURE数据集主要测试模型对头部运动的鲁棒性,如表1所示,TAST-Net表现出了最优的性能,取得了最小的e_{RMS}(0.53%)和e_{MA}(0.37%),以及最大的R值(0.96).这表明TAST-Net能够以非常高的精度和相关性估计SpO_2,并且在处理PURE数据集中的运动干扰方面具有出色的鲁棒性.MMFM(e_{RMS}=0.89%,e_{MA}=0.66%,R=0.87)表现次优.值得注意的是,经典的3D-CNN模型在此数据集上的表现不佳(e_{RMS}=6.13%,e_{MA}=5.89%,R=0.25),这可能反映了其在没有特定增强机制的情况下,处理rPPG信号中的运动噪声时面临的挑战.

VIPL-HR数据集包含更大规模、更复杂的环境场景.如表2所示,TAST-Net性能最优(e_{RMS}=0.84%,e_{MA}=0.57%,R=0.82);MultiPhysNet(e_{RMS}=0.90%,e_{MA}=0.66%,R=0.80)表现次优.这一结果凸显了TAST-Net在应对光照变化、不同头部姿态等真实环境场景挑战时,依然能够保持强大的泛化能力和估计精度.其余对比方法的性能则出现了不同程度的下降,例如MMFM模型在该数据集上的R值相对较低(0.59),表明其融合策略在复杂场景下的适应性有待提升.

从模型效率的角度分析,各方法展现了不同的设计权衡.MMFM与MultiPhysNet是典型的轻量级模型,参数量均在1×10^6左右,其中MMFM推理速度最快(0.077 ms/帧),但其在复杂场景下的精度有所牺牲.与之相反,基准3D-CNN模型因其未经优化的网络结构,计算效率最低(1.182 ms/帧).本研究提出的TAST-Net则在性能与效率间取得了显著平衡,其参数量(3.24×10^6)适中,推理速度(0.149 ms/帧)极具竞争力,在实现最优估计精度的同时,也展现了高效的计算性能,证明了其架构设计的优越性.

为了更直观地评估TAST-Net模型的估计一致性与个体样本的偏差情况,本研究还对其估计结果进行了可视化分析.图3展示了TAST-Net在2个公开数据集上的Bland-Altman图和散点图.

图3中可视化结果进一步证实了TAST-Net的性能优越性.Bland-Altman结果显示,在PURE数据集上,TAST-Net估计值与真实值的平均偏差仅为0.16%,且95%的一致性界限(limits of agreement)位于-0.84%~1.16%这一狭窄区间内,表明2种估计结果具有良好的一致性.而在更具挑战性的VIPL-HR数据集上,该模型依然表现稳健,平均偏差为0.06%,95%的一致性界限为-1.58%~1.71%.根据国际标准ISO 80601-2-61[30]对医用脉搏血氧仪的要求,其e_{RMS}需小于3%.本研究中TAST-Net在VIPL-HR上的e_{RMS}=0.84%,远低于该临床标准.这些具体的定量指标表明,TAST-Net的估计结果不仅系统性偏差极小,而且绝大多数估计误差都在临床可接受的范围内,从而在统计学上验证了其估计结果的准确性和可靠性.

2.8 消融实验

为全面验证TAST-Net中各关键组成部分的有效性,本研究设计了一系列消融实验.实验不仅探究模型核心组件(即双路融合架构与组合损失函数)的贡献,还量化关键预处理步骤(即欧拉视频放大)对最终性能的实际影响.首先,为评估模型核心组件的作用,在PURE数据集上进行了实验,具体设置如下:

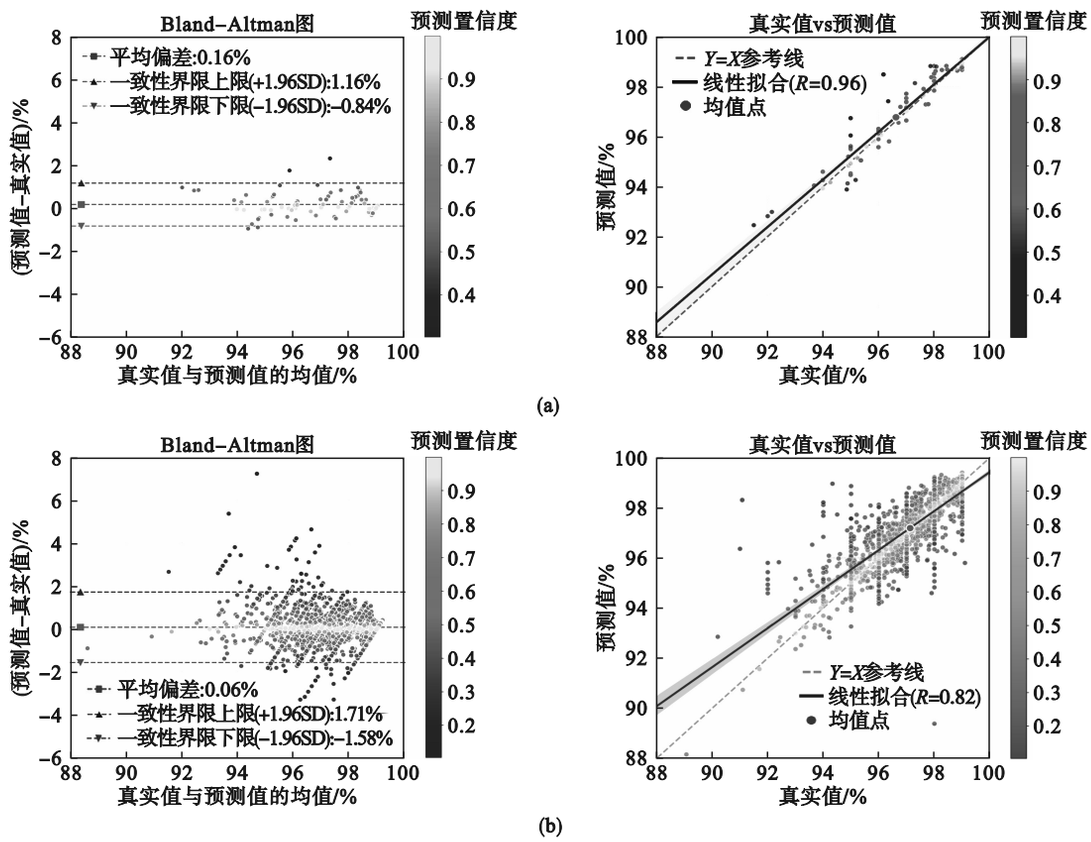
1) Baseline:仅使用ViViT单分支模型进行SpO_2估计,并采用标准的MSE损失函数.

2) Baseline+Dual-Path Architecture:在Baseline的基础上,加入3D CNN分支构成双路融合架构,但仍然使用标准的MSE损失函数进行训练.此项实验旨在独立验证双路融合架构本身带来的性能增益.

3) Baseline+Total Loss(L_T):在ViViT单分支模型的基础上,将损失函数替换为本文提出的组

合损失函数(L_T).此项实验旨在独立验证组合损失函数的有效性.

4) TAST-Net:即本文提出的完整模型,采用ViViT+3D CNN的双路融合架构,并使用组合损失(L_T)函数进行训练.



注:左侧为 $SpO_{2_{pre}}$ 和 $SpO_{2_{gt}}$ 的Bland-Altman图;右侧为 $SpO_{2_{pre}}$ 和 $SpO_{2_{gt}}$ 的散点图;SD(standard deviation),±1.96SD表示95%的一致性界限.

图3　TAST-Net在公开数据集上的血氧饱和度估计结果可视化

Fig. 3　Visualization of blood oxygen saturation estimation results of TAST-Net on public dataset

(a)—PURE数据集;(b)—VIPL-HR数据集.

消融实验的结果如表3所示.

表3　TAST-Net模型消融实验结果

Table 3　Ablation experiment results of TAST-Net model

模型	e_{RMS}/%	e_{MA}/%	R
Baseline	1. 32	1. 05	0. 68
Baseline+Dual-Path Architecture	0. 58	0. 41	0. 95
Baseline+Total Loss(L_T)	1. 12	0. 85	0. 84
Our TAST-Net	**0. 53**	**0. 37**	**0. 96**

如表3所示,消融实验的结果量化了TAST-Net各核心组件的贡献.

首先,双路融合架构的引入对模型性能有决定性影响.与仅使用ViViT的基准(Baseline)模型相比,采用双路融合架构但仍使用MSE损失的模型(Baseline+Dual-Path Architecture)的性能得到了全面提升,其 e_{RMS} 从1.32%显著降低至0.58%,e_{MA} 从1.05%降低至0.41%,同时 R 从0.68大幅提升至0.95.这一结果表明,通过结合3D CNN对局部生理细节的捕捉能力和ViViT对全局时空依赖的建模能力,是提升模型估计精度的核心因素.

其次,组合损失函数的有效性也得到了验证.在基准ViViT模型上仅将损失函数替换为组合损失(L_T)函数后,模型的 e_{RMS} 从1.32%降低至1.12%,e_{MA} 从1.05%降低至0.85%,R 值则从0.68显著提升至0.84.这证明通过优化趋势相关性可以有效改善模型的估计可靠性.

最终,完整的TAST-Net模型(结合了双路融合架构与组合损失)在PURE数据集上取得了最优性能(e_{RMS}=0.53%,e_{MA}=0.37%,R=0.96).该性能优于任何单独引入组件的模型.值得注意的是,与仅采用双路融合架构的模型(e_{RMS}=0.58%,e_{MA}=0.41%,R=0.95)相比,完整的TAST-Net通过结合

组合损失函数，进一步将 e_{RMS} 降低至 0.53%，将 e_{MA} 降低至 0.37%，并将 R 值提升至 0.96. 这一增量改进清晰地表明，本文提出的双路融合架构与组合损失函数之间存在有效的协同作用，二者结合能够最大化提升模型 SpO_2 估计的性能.

除模型自身组件外，前端的数据预处理对性能同样至关重要. 为此，实验进一步对欧拉视频放大(EVM)预处理步骤的有效性进行了验证. 本文在 PURE 和 VIPL-HR 数据集上，使用“有 EVM”和“无 EVM”两种预处理方式的数据，分别训练和测试了 TAST-Net 模型，结果如表 4 所示.

表 4 EVM 预处理消融实验结果

Table 4 Ablation experiment results of EVM preprocessing

数据集	预处理方式	e_{RMS}/%	e_{MA}/%	R
PURE	TAST-Net(无 EVM)	0.59	0.46	0.94
	TAST-Net(有 EVM)	**0.53**	**0.37**	**0.96**
VIPL-HR	TAST-Net(无 EVM)	0.91	0.66	0.79
	TAST-Net(有 EVM)	**0.84**	**0.57**	**0.82**

从表 4 的对比结果可以清晰地看出，EVM 预处理步骤对 TAST-Net 模型的性能有显著的提升作用. 在 2 个数据集上，未使用 EVM 进行预处理的模型，其各项误差指标均明显升高，相关系数则出现显著下降. 这有力地证明了 EVM 作为一种信号前置增强技术，通过选择性地放大与心率频带匹配的微弱颜色变化，有效提升了输入视频中原始 rPPG 信号的信噪比(SNR). 消融实验验证了 EVM 是整体方法框架中不可或缺的一环，它与 TAST-Net 的深度特征提取能力形成了有效互补，共同确保了最终 SpO_2 估计的准确性与鲁棒性.

3 结 论

1) 提出了趋势感知时空融合网络模型 TAST-Net. 其采用的 3D CNN 与 ViViT 双路融合架构，能够有效协同处理面部视频中的局部生理细节与全局时空依赖性，克服了单一模型在时空特征提取上的局限性.

2) 在 PURE 和 VIPL-HR 两个公开数据集上的综合实验结果表明，TAST-Net 在均方根误差(e_{RMS})、平均绝对误差(e_{MA})和皮尔逊相关系数(R)等关键性能指标上，均优于多种对比的深度学习模型，证明了所提方法的优越性与良好的泛化能力.

3) 通过创新的网络结构和针对性的损失函数设计，TAST-Net 在实现高精度的同时，也展现了高效的计算性能，为从面部视频中进行精准、稳健的非接触式 SpO_2 估计提供了一个有效的解决方案.

参考文献：

[1] Laratta C R, Ayas N T, Povitz M, et al. Diagnosis and treatment of obstructive sleep apnea in adults[J]. *Canadian Medical Association Journal*, 2017, 189(48): 1481-1488.

[2] Watson A R, Wah R, Thamman R. The value of remote monitoring for the COVID-19 pandemic[J]. *Telemedicine Journal and e-Health*, 2020, 26(9): 1110-1112.

[3] Amoore J N. Pulse oximetry: an equipment management perspective[C]//IEE Colloquium on Pulse Oximetry: A Critical Appraisal. London, 2002: 124-126.

[4] Shimazaki T, Hara S, Okuhata H, et al. Cancellation of motion artifact induced by exercise for PPG-based heart rate sensing[C]// The 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Chicago, 2014: 3216-3219.

[5] Verkruysse W, Svaasand L O, Nelson J S. Remote plethysmographic imaging using ambient light[J]. *Optics Express*, 2008, 16(26): 21434-21445.

[6] de Haan G, Jeanne V. Robust pulse rate from chrominance-based rPPG [J]. *IEEE Transactions on Bio-medical Engineering*, 2013, 60(10): 2878-2886.

[7] Poh M Z, McDuff D J, Picard R W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation[J]. *Optics Express*, 2010, 18(10): 10762-10774.

[8] Balakrishnan G, Durand F, Guttag J. Detecting pulse from head motions in video [C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, 2013: 3430-3437.

[9] Wang W J, den Brinker A C, Stuijk S, et al. Algorithmic principles of remote PPG [J]. *IEEE Transactions on Biomedical Engineering*, 2017, 64(7): 1479-1491.

[10] Chen W X, McDuff D. DeepPhys: video-based physiological measurement using convolutional attention networks [C]//Computer Vision-ECCV 2018. Cham: Springer, 2018: 356-373.

[11] Mathew J, Tian X, Wong C W, et al. Remote blood oxygen estimation from videos using neural networks[J]. *IEEE Journal of Biomedical and Health Informatics*, 2023, 27(8): 3710-3720.

[12] Yu Z T, Shen Y M, Shi J G, et al. PhysFormer++: facial video-based physiological measurement with slow fast temporal difference transformer[J]. *International Journal of Computer Vision*, 2023, 131(6): 1307-1330.

[13] Yu Z T, Shen Y M, Shi J G, et al. PhysFormer: facial video-based physiological measurement with temporal difference transformer [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, 2022: 4176-4186.

[14] Du J D, Liu S Q, Zhang B C, et al. Weakly supervised rPPG estimation for respiratory rate estimation[C]// IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, 2021: 2391-2397.

[15] Gideon J, Stent S. The way to my heart is through contrastive learning: remote photoplethysmography from unlabelled video [C]//IEEE/CVF International Conference

on Computer Vision (ICCV). Montreal, 2022: 3975-3984.
[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. Long Beach ,CA,2017:5998-6008.
[17] Arnab A, Dehghani M, Heigold G, et al. ViViT: a video vision transformer [C]// IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, 2021: 6816-6826.
[18] Stricker R, Müller S, Gross H M. Non-contact video-based pulse rate measurement on a mobile service robot[C]//The 23rd IEEE International Symposium on Robot and Human Interactive Communication. Edinburgh, 2014: 1056-1062.
[19] Niu X S, Han H, Shan S G, et al. VIPL-HR: a multi-modal database for pulse estimation from less-constrained face video [C]//Computer Vision-ACCV 2018. Cham: Springer, 2018: 562-576.
[20] Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees[C]// IEEE Conference on Computer Vision and Pattern Recognition. Columbus, 2014: 1867-1874.
[21] Lugaresi C, Tang J Q, Nash H, et al. MediaPipe: a framework for building perception pipelines [EB/OL]. (2019-06-12) [2024-11-19]. https://arxiv. org/abs/1906.08172.
[22] Wu H Y, Rubinstein M, Shih E, et al. Eulerian video magnification for revealing subtle changes in the world[J]. *ACM Transactions on Graphics*, 2012, 31(4): 1-8.
[23] Loshchilov I, Hutter F. Decoupled weight decay regularization [EB/OL]. (2017-11-14) [2024-12-25]. https://arxiv.org/abs/1711.05101.
[24] He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]// IEEE International Conference on Computer Vision (ICCV). Santiago, 2015: 1026-1034.
[25] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, 2016: 770-778.
[26] Kim S Y, Lim J, Na T, et al. 3DSRnet: video super-resolution using 3D convolutional neural networks [EB/OL]. (2018-12-21) [2024-11-15]. https://arxiv. org/abs/1812.09079.
[27] Liu K, Tang J K, Jiang Z, et al. Summit vitals: multi-camera and multi-signal biosensing at high altitudes[C]// IEEE Smart World Congress (SWC). Nadi, 2025: 284-291.
[28] Zhu S W, Liu S H, Jing X J, et al. Innovative approaches in imaging photoplethysmography for remote blood oxygen monitoring[J]. *Scientific Reports*, 2024, 14: 19144.
[29] Hu M, Wu X, Wang X H, et al. Contactless blood oxygen estimation from face videos: a multi-model fusion method based on deep learning[J]. *Biomedical Signal Processing and Control*, 2023, 81: 104487.
[30] Respiratory Devices and Related Equipment Used for Patient Care. Medical electrical equipment. Part 2-61: particular requirements for basic safety and essential performance of pulse oximeter equipment: ISO 80601-2-61:2017[S/OL]. (2017-12-15)[2025-03-12]. https://ftp.kssn.net/for/fordetail.do? itemNo=F011010084595.

(上接第 19 页)

[21] Zabala A, Pons X. Effects of lossy compression on remote sensing image classification of forest areas[J]. *International Journal of Applied Earth Observation and Geoinformation*, 2011, 13(1): 43-51.
[22] Chao J S, Steinbach E. Preserving SIFT features in JPEG-encoded images [C]//The 18th IEEE International Conference on Image Processing. Brussels, 2011: 301-304.
[23] Shindo T, Yamada K, Watanabe T, et al. Image coding for machines with edge information learning using segment anything [C]// IEEE International Conference on Image Processing (ICIP). Abu Dhabi, 2024: 3702-3708.
[24] Paniga S, Borsani L, Redondi A, et al. Experimental evaluation of a video streaming system for wireless multimedia sensor networks [C]// The 10th IFIP Annual Mediterranean Ad Hoc Networking Workshop. Favignana Island, 2011: 165-170.
[25] Bernal J, Sánchez J, Vilariño F. Towards automatic polyp detection with a polyp appearance model [J]. *Pattern Recognition*, 2012, 45(9): 3166-3182.
[26] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]// IEEE International Conference on Computer Vision. Venice, 2017: 2980-2988.
[27] Liu J Y, Lian J, Yu Y Z. ChestX-Det10: chest X-ray dataset on detection of thoracic abnormalities [EB/OL]. (2020-06-17) [2025-05-10]. https://arxiv. org/abs/2006.10550.
[28] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137-1149.
[29] Minnen D, Ballé J, Toderici G. Joint autoregressive and hierarchical priors for learned image compression[EB/OL]. (2018-09-08) [2025-05-10]. https://arxiv. org/abs/1809.02736.
[30] Cheng Z X, Sun H M, Takeuchi M, et al. Learned image compression with discretized Gaussian mixture likelihoods and attention modules[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 7936-7945.
[31] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 3213-3223.
[32] Romera E, Álvarez J M, Bergasa L M, et al. ERFNet: efficient residual factorized ConvNet for real-time semantic segmentation [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2018, 19(1): 263-272.