doi:10.3772/j.issn.1002 – 0470.2026.08.002

融合部分-整体时间窗口及双注意力机制的视频心率提取网络①

赵昶辰②[*,**] 曹鹏程[*] 黄 斌[***] 冯远静[*]

([*]浙江工业大学信息工程学院 杭州 310023)

([**]杭州电子科技大学计算机学院 杭州 310018)

([***]北京航空航天大学杭州创新研究院 杭州 310051)

摘 要 虽然近年来出现了许多远程光体积描记(remote photoplethysmography,rPPG)心率提取神经网络,但是大多只考虑了时空连续性而忽略了不可预测的视频数据异常情况,导致心率识别精准度下降。为此本文提出了一种基于部分-整体时间窗口特征融合与时间-通道联合注意力机制的视频心率提取网络,旨在当视频中出现异常情况时仍能够稳定、准确地估计心率,并将注意力机制与损失函数相结合定义联合损失函数。所提网络在PURE和UBFC-rPPG这2个数据集上与现有的9种方法进行比较,以验证所提模型的预测能力。实验结果表明,在PURE数据集上,相比于次优的模型,所提方法其信噪比(signal-to-noise ratio,SNR)提高了25.21%,平均绝对误差(mean absolute error,MAE)降低了31.63%,均方根误差(root mean square error,RMSE)降低了34.00%,ρ提高了4.25%;在UBFC-rPPG数据集上,相比于次优的模型,所提方法其SNR提高了13.62%,MAE降低了15.38%,RMSE降低了47.05%,ρ提高了0.40%。实验证明融合部分-整体时间窗口特征以及时间-通道联合注意力机制可有效抑制噪声干扰,进一步提升心率测量精度。

关键词 远程光体积描记;神经网络;心率估计;分离时间窗口;视频处理

心率是反映人体健康状况的最重要生命体征之一,传统心率测量[1]方法常使用心电图或光体积描记(photoplethysmography,PPG)。然而,这些方法需与人体皮肤直接接触,容易产生交叉感染并带来清洁维护成本,不适用于不方便佩戴的场合。远程光体积描记(remote photoplethysmography,rPPG)实现了从接触式测量到非接触式测量的技术飞跃,已成为近年来十分活跃的研究课题[2]。rPPG使用消费级数码相机,从视频[3]中检测由心脏周期性活动引起的皮肤颜色周期性变化,进而重建血液容积脉搏波并进行生理指标的测量。与传统接触式PPG测量相比,rPPG不需要与受试者身体进行任何接触,特别适合于难以佩戴传感器的场合。例如,烧伤患者监测[4]、驾驶员检测[5]、新生儿监测[6]、睡眠监测[7]、健身运动[2]。在视频数据中,脉搏信号常常包含各种噪声,包括光照强度的变化、人脸的运动、相机量化噪声等。因此,去除观测信号中的噪声信号、提高识别精度一直是rPPG脉搏提取研究的首要问题。

早期的rPPG测量技术基于传统方法,一般分为3个阶段。首先,通过目标跟踪算法从视频中获取稳定的人脸区域,之后通过平均颜色通道内的像素值获得三通道原始测量信号。然后,采用信号分离模型、相机成像模型等模型从原始测量信号中提取脉搏信号。最后,根据脉搏信号计算心率。其中,第2阶段是核心,模型包括基于盲源分离的方

① 国家自然科学基金面上项目(62573171)和国家自然科学基金青年基金(62203037)资助项目。

② 男,1988年生,博士,副教授,研究方向:计算机视觉,视觉生理信号测量与人工智能;联系人,E-mail:cczhao@zjut.edu.cn。

(收稿日期:2024 – 05 – 07)

法[8-9],基于成像模型的方法[10-11]以及数据驱动的方法[12]等。此外,人脸检测与跟踪算法[13]和视频压缩[14]的不稳定性也是对 rPPG 测量精度有显著影响的重要因素。随着深度学习的飞速发展,神经网络也被广泛应用到 rPPG 检测中,例如 Spetlik 等人[15]提出的 HR-CNN(heart rate- convolutional neural network)方法,Chen 和 McDuff[16]提出的 DeepPhys,以及基于 DeepPhys 改进的 MTTS-CAN(multi-task temporal shift convolutional attention network)[17]方法。由于二维(two-dimensional,2D)卷积只考虑了视频帧的空间信息而对时域关联性建模不足,研究者们将时空网络引入 rPPG 模型,这些时空网络可以更有效地表示视频流中生理信号的时空信息,如 PhysNet[14]、ETA-rPPGNet[18]、DeeprPPG[19]和时域差分卷积[20]。

无论是使用常规的 2D 卷积还是三维(three-dimensional,3D)时空卷积的神经网络模型,对于输入视频中存在数据异常的情况仍会导致网络性能不佳。在实际应用过程中,被拍摄者很容易会出现突然转头、头部移出画面、被遮挡、跟踪框跳变或消失等情况,这些情况会导致视频数据中出现信号极速跳变、部分帧信号缺失等情形。现有大部分神经网络并没有对此类情况的应对策略,而将这些干扰的、无效的信息看作学习过程中的有效信息,导致神经网络性能降低。解决这一问题的基本思想是降低这些含有干扰无效信息在训练数据中的比重,从而减少其对整个深度学习过程的影响。然而问题的难点在于,干扰信息的出现是不规律的且难以被预测,在学习前的数据预处理环节也很难用简单的手段进行剔除[21],这使得问题更具挑战性。

为了解决这一问题,本文提出一种融合部分-整体时间窗口及双注意力机制的神经网络,旨在能够自适应视频中的异常情况,可以稳定输出人体心率,保持较高的测量精度。

1 相关工作

1.1 远程视频心率提取

光体积描记于 20 世纪 30 年代[22]提出,并因其便利性和无创性被广泛使用。在 PPG 的基础上,研究者们发现利用光学相机也可测量生理信号,例如,心率和呼吸率。由于该方法原理与 PPG 类似,且可以远程测量,因此人们将该技术称之为远程光体积描记。

Haan 和 Jeanne[10]提出了基于色度模型的算法 CHROM(chrominance),他们使用色度来消除相机捕获的图像中的镜面反射分量。同时,他们提出进行肤色标准化,以使算法适应不同颜色的光源。Wang 等人[11]进一步探索了该算法,并提出 POS(plane orthogonal to skin)算法,该方法是基于皮肤反射模型,利用血容量脉搏的知识在与时间归一化肤色方向正交的平面上定义一个粗略的投影区域,并通过实时调优获得在平面上精确的投影方向,通过不断地迭代优化输出最终的心率信号,该算法削弱了 CHROM 中标准化肤色的必要性。

近年来,卷积神经网络(convolutional neural networks,CNNs)被广泛引入到 rPPG 信号测量中。Spetlik 等人[15]提出了一种两阶段的 HR-CNN 方法,它包括一个测量 rPPG 信号的提取器和一个获得心率的估计器。DeepPhys[16]是一个端到端系统,使用深度卷积网络基于双路 CNN 网络来测量心率和呼吸速率。Yu 等人[14]设计了一个时空网络 PhysNet,在处理空间信息时同时考虑时间信息的影响。Liu 等人[17]提出了一种时移卷积注意网络结构 MTTS-CAN,通过沿时间维度移动张量来促进跨多帧的信息交互。Yu 等人[20]提出了时差卷积运算来捕捉帧间关系。因上述大部分神经网络对输入特征的整体时间维度未进行处理,进行特征信息感知时往往无法分辨数据真实性,因而对存在异常数据信息的数据往往性能效果较差。

本文为减少这种影响,提出一种部分-整体时间窗口相结合的网络结构,将时间维度分割为若干时间窗口分别处理,降低时域无效信号的权重,进而减少对学习过程的影响,突出有效时域信号的权重,使脉搏信号特征更易获取。这样做的优点是异常数据会落入到某些子窗口中而不会对整个训练网络产生较大的影响。由于异常数据在整体训练数据中的比重很小,算法自然不会为这类特征分配很高权重,这

样即在算法层面实现了自动抑制异常数据的目的。

1.2 注意力机制

受人类视觉系统的启发,神经网络中的注意力机制可以有效突出特定物体或场景的特定部分。注意力机制已成功地应用于场景分割[23]、视频识别[24]、机器翻译[25]、视觉同时定位与地图构建[26]等多个领域。Kong 等人[27]提出了用于动作识别的对称增强融合网络。该网络由 2 个流组成,空间流通过图卷积网络从骨骼中提取空间特征,时间流利用嵌入式运动序列计算算法从骨骼数据中提取时间特征。Zhang 等人[28]提出了用于视频去模糊的多注意卷积神经网络,使用一种时空注意力模块来捕捉相邻帧间存在的时空信息的特征。在 rPPG 领域,注意力机制也被证明是提取脉冲信号的有效方法。DeepPhys[16]从外观表征中获得空间注意图来指导特征在运动表示法中的提取。Hu 等人[18]提出一种时域注意力网络 ETA-rPPGNet,该网络利用一维卷积有效地建模局部时域的信息关联。Efficient-Phys[29]提出了一个自注意力转移网络(self-attention-shifted network,SASN),它建立在 TS-CAN(temporal-shift convolutional attention network)中的时空建模方法之上,具有 2 个卷积分支,其中一个采用预处理的差异帧表示,另一个采用归一化外观帧。

本文所提的注意力机制为双注意力机制,即同时学习一个时间注意力权重和一个通道注意力权重。所提双注意力机制与现有注意力机制有以下 3 个方面不同。(1)与将空间特征和通道特征作为一个整体进行平均处理的一维时域注意相比,所提注意力机制实现了时间和通道维度的精细化注意力学习。(2)虽然现有方法[21]在空间-时间和通道-时间的联合空间中学习注意力权重的模式能够更加准确地描述脉搏在联合空间中的传播特性,但是其计算量巨大。本文在保证双注意力权重学习的前提下节约计算消耗,为部分-整体时间窗口融合的实现提供可能。(3)现有注意力机制往往关注于对网络训练过程中某一层的注意力,而本文的注意力机制通过输入数据产生相关特征信号参与了网络的损失函数构建。

2 所提算法

2.1 网络总体结构

所提 rPPG 脉搏提取网络的总体结构如图 1 所示。输入是视频序列,输出是脉搏信号。算法由分离时间-通道联合注意力网络模块(separated time-channel joint attention network, STCJANet)和 3D 卷积模块和池化层组成。卷积模块与池化层负责提取视频中与 rPPG 信号相关的特征,STCJANet 负责加强所提特征中与 rPPG 信号有关的模式。每个卷积模块包含一个 3D 卷积层、ReLU 层与批量归一化层。本算法除了在网络最终输出处定义一个损失函数外,在每个 STCJANet 处也定义了一个损失项,与最终的输出损失项共同构成损失函数,旨在加速网络收敛。

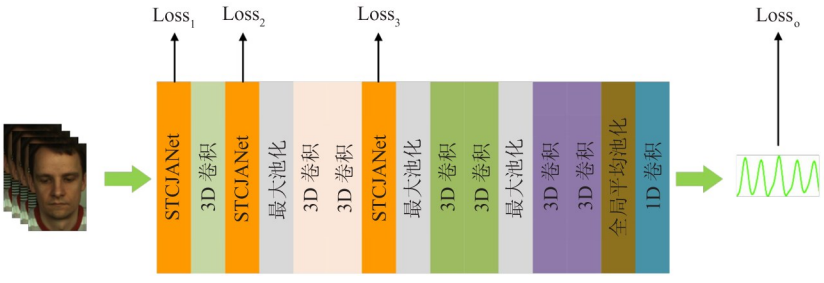

图 1 rPPG 信号提取网络结构图

STCJANet 包含分离时间窗口网络模块(separated time window network,STWNet)和全局时域网络模块(global time domain network, GTNet)。其中 STWNet 将输入特征在时间维度进行分割,利用异常数据在整体训练数据中占比较少这一先验经验,学习时间注意力权重和通道注意力权重,以实现弱化无效信息对网络性能的影响并突出重要脉搏信号特征。GTNet 旨在为了弥补分离时间窗口带来的时域

不连续性，以感知连续时间特征以及全局时间特征。

表 1 描述一个基于 STCJANet 的神经网络架构。表中展示了每一层的卷积滤波器大小和输出特征图尺寸，未显示注意力模块。输入数据的窗口长度 T 根据实验设置。

表 1　rPPG 脉搏提取网络架构规范

网络层	滤波器尺寸	输出尺寸
输入数据	无	$3 \times T \times 192 \times 128$
Conv3d-1	$3 \times 1 \times 5 \times 5$	$3 \times T \times 192 \times 128$
MaxPool3d-1	$1 \times 2 \times 2$	$3 \times T \times 96 \times 64$
Conv3d-2-1	$3 \times 3 \times 3 \times 3$	$3 \times T \times 96 \times 64$
Conv3d-2-2	$3 \times 3 \times 3 \times 3$	$3 \times T \times 96 \times 64$
MaxPool3d-2	$1 \times 2 \times 2$	$3 \times T \times 48 \times 32$
Conv3d-3-1	$32 \times 3 \times 3 \times 3$	$32 \times T \times 48 \times 32$
Conv3d-3-2	$32 \times 3 \times 3 \times 3$	$32 \times T \times 48 \times 32$
MaxPool3d-3	$1 \times 2 \times 2$	$32 \times T \times 24 \times 16$
Conv3d-4-1	$64 \times 3 \times 3 \times 3$	$64 \times T \times 24 \times 16$
Conv3d-4-2	$64 \times 3 \times 3 \times 3$	$64 \times T \times 24 \times 16$
AvgPool	24×16	$64 \times T \times 1 \times 1$
Conv1d-last	$1 \times 1 \times 1 \times 1$	$1 \times T \times 1 \times 1$

第 1 个 3D 卷积核的大小是 $1 \times 5 \times 5$，步幅为 1，填充为 $0 \times 2 \times 2$ 以保证输出特征图尺寸保持与输入相同的高宽比，其他的 3D 卷积滤波器大小设置为 $3 \times 3 \times 3$，步幅为 1，以保持输出大小不变。通道数由红绿蓝（red，green，blue，RGB）三通道先扩展到了 32，随后拓展到了 64。最大池化层的内核大小为 $1 \times 2 \times 2$，步幅为 $1 \times 2 \times 2$，使得输出尺寸减半。

2.2　分离时间与通道联合注意力模块

训练数据的异常情况，如某帧出现跟踪框抖动、人脸移出拍摄范围、人脸突然转动等，导致该帧图像数据与前后帧不连续，由于 3D 卷积关注时空的连续性而忽略了异常信号的跳变性，导致最终的识别精度受到影响。

针对这个问题，本文设计了分离时间与通道联合注意网络模块 STCJANet，能够有效地对输入数据集的时间与通道两维进行权重重置，有效提高了网络的抗干扰能力。该模块体现 2 处融合，即部分时间窗口与全局时间窗口的结合和时间注意力权重与通道注意力权重的融合。图 2 展示了 STCJANet 的网络结构，包括分离时间窗口网络模块和全局时域网络模块。

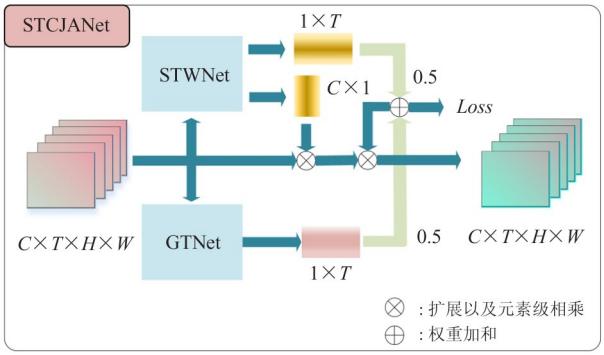

图 2　所提分离时间与通道联合注意力网络模块的结构图

STWNet 对特征图 $\boldsymbol{x} \in \mathbb{R}^{C\times T\times H\times W}$ 在时间维度先分割后进行特征提取，输出一个时域注意力变量 $\boldsymbol{a}_{\mathrm{stw}}^{t} \in \mathbb{R}^{T}$ 和一个通道注意力变量 $\boldsymbol{a}_{\mathrm{stw}}^{c} \in \mathbb{R}^{C}$。其中，$C$ 表示通道数，T 表示时间维度，H 和 W 表示输入图片的二维尺寸。GTNet 则对输入特征不进行时间维度分割，输出一个时域注意力变量 $\boldsymbol{a}_{\mathrm{gt}}^{t} \in \mathbb{R}^{T}$。总体的时域注意力变量为 2 个子网络时域注意力变量的加权和 $\boldsymbol{a}^{t}$。

$$\boldsymbol{a}^{t} = \omega_1 \cdot \boldsymbol{a}_{\mathrm{stw}}^{t} + \omega_2 \cdot \boldsymbol{a}_{\mathrm{gt}}^{t} \tag{1}$$

为了能够得到 rPPG 相关的注意力信息，此处计算损失函数，通过引入真值特征对 $\boldsymbol{a}^{t}$ 进行监督。整个网络最终的输出为输入特征图经过通道注意力变量与时域注意力变量作用后的特征图 $\tilde{\boldsymbol{x}}$。

$$\tilde{\boldsymbol{x}} = Attn(Attn(\boldsymbol{x} \mid \boldsymbol{a}_{\mathrm{stw}}^{c})\boldsymbol{a}^{t}) \tag{2}$$

式中：$Attn(\cdot \mid \boldsymbol{u})$ 为以 $\boldsymbol{u}$ 为注意力变量的扩展以及元素级相乘[30]操作。

2.3　STWNet 网络模块

STWNet 网络模块的总体结构如图 3 所示。对于输入特征图 $\hat{\boldsymbol{x}} \in \mathbb{R}^{C\times T\times H\times W}$ 在时间维度分离为 N 个不交叠的窗口，随后分别输入分离时间窗口模块进行处理，输出 N 个尺寸为 $1 \times T$ 和 $C \times 1$ 的注意力变量，其中 $t = \dfrac{T}{N}$。再将这些注意力变量按照原始窗口顺序进行拼接（concatenation），即得到了最终的时域注意力变量和通道注意力变量。

对于每个分离时间窗口的数据 $\boldsymbol{x} \in \mathbb{R}^{C\times t\times H\times W}$，分 3 条通路进行处理，分别为时间注意力通路、通道注意力通路和时间-通道联合注意力通路。每条通

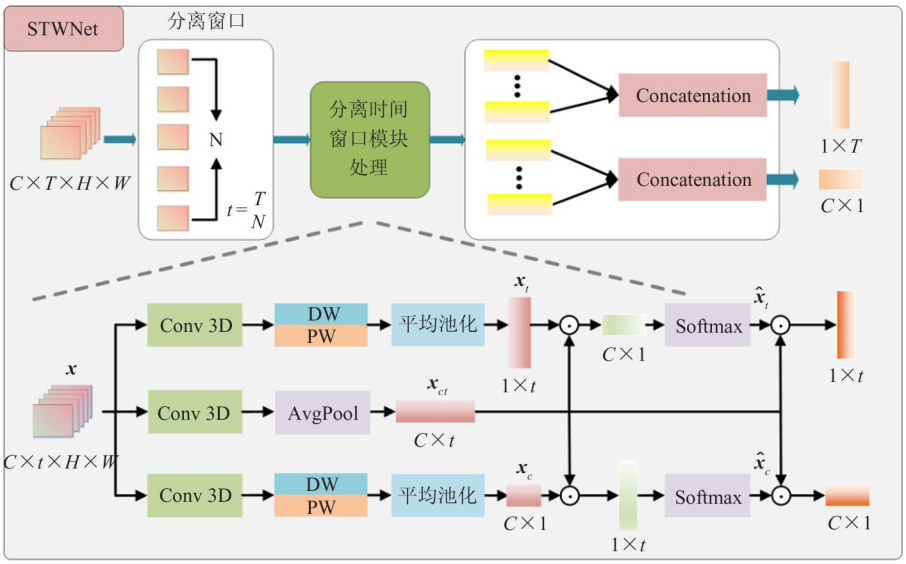

图 3　所提分离时间窗口网络模块的结构图

路均首先对输入数据进行一次 3D 卷积。对于时间-通道联合注意力通路,将特征图的空间维度进行平均池化(AvgPool),得到时间-通道联合注意力变量 $\boldsymbol{x}_{ct}\in\mathbb{R}^{C\times t}$。对于时间注意力通路与通道注意力通路,则分别进行深度可分离卷积(DW/PW)与池化操作,得到通道注意力变量 $\boldsymbol{x}_c\in\mathbb{R}^{C}$ 与时域注意力变量 $\boldsymbol{x}_t\in\mathbb{R}^{t}$。然后,将时间注意力变量与时间-通道联合注意力变量按照式(3)与式(4)操作,得到最终的输出时间注意力变量 $\boldsymbol{y}_t\in\mathbb{R}^{t}$。

$$\hat{\boldsymbol{x}}_t=\mathrm{Softmax}(\boldsymbol{x}_{ct}\odot\boldsymbol{x}_t)\tag{3}$$

$$\boldsymbol{y}_t=\boldsymbol{x}_{ct}^{\mathrm{T}}\odot\hat{\boldsymbol{x}}_t\tag{4}$$

式中:Softmax 函数用于多分类过程,将多个神经元的输出映射到(0,1)区间;⊙表示矩阵相乘;上标 T 表示转置。类似地,将通道注意力变量按照式(5)与式(6)进行操作,得到输出的通道注意力变量 $\boldsymbol{y}_c\in\mathbb{R}^{C}$。

$$\hat{\boldsymbol{x}}_c=\mathrm{Softmax}(\boldsymbol{x}_{ct}^{\mathrm{T}}\odot\boldsymbol{x}_c)\tag{5}$$

$$\boldsymbol{y}_c=\boldsymbol{x}_{ct}\odot\hat{\boldsymbol{x}}_c\tag{6}$$

2.4　GTNet 网络模块

GTNet 网络模块的总体结构如图 4 所示,旨在通过全局时域网络模块以弥补分离时间窗口带来的时域不连续性。首先将输入特征图在 *H-T*、*H-W*、*W-T* 这 3 个角度分别进行二维卷积,旨在通过不同的二维维度以感知不同维度间的关联有效信息,之后进行 6 次膨胀率(dilated rate)分别为 1、2、3、1、2、3 的膨胀卷积(dilated Conv),以扩大感知视野,获得更多信息,之后沿通道维度进行拼接,获得维度大小为 $3C\times T\times H\times W$ 的数据,最后进行一维卷积和池化,得到 $1\times T$ 的时域注意力变量。

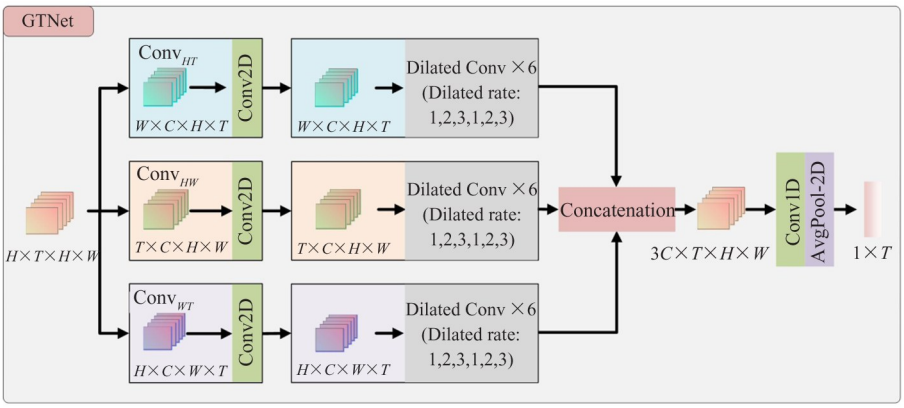

图 4　所提全局时域网络模块的结构

2.5 损失函数

所提网络的损失函数在每个 STCJANet 网络和最终的输出上均有定义。如图 1 所示,网络中 3 个 STCJANet 网络分别定义一个损失函数,表示为 Loss_1、Loss_2、Loss_3。网络输出的损失函数表示为 Loss_o。上述损失函数均采用负皮尔逊相关系数作为损失函数,设 $\boldsymbol{x} = [x_1, x_2, \cdots, x_T]^{\mathrm{T}}$, $\boldsymbol{y} = [y_1, y_2, \cdots, y_T]^{\mathrm{T}}$ 分别为网络输出的脉冲信号和真实的 PPG 信号,其中, T 为信号长度。网络总体的损失函数为

$$\mathrm{Loss} = \alpha \cdot \mathrm{Loss}_1 + \beta \cdot \mathrm{Loss}_2 + \theta \cdot \mathrm{Loss}_3 + \mathrm{Loss}_o \tag{7}$$

式中: α、β、θ 为权重系数。Loss 采用负皮尔逊相关系数损失函数,计算方法如下:

$$\mathrm{Loss}_{\mathrm{pearson}} = 1 - \frac{N\sum_1^N x_i y_i - \sum_1^N x_i \sum_1^N y_i}{\sqrt{(N\sum_1^N x_i^2 - (\sum_1^N x_i)^2)(N\sum_1^N y_i^2 - (\sum_1^N y_i)^2)}} \tag{8}$$

3 实验设置

3.1 数据集

本文使用 2 个公开数据集验证算法性能,分别为 PURE 数据集[3]和 UBFC-rPPG 数据集[31]。

PURE:该数据集包含 59 段视频,10 名健康受试者进行 6 种头部运动,包括静止、讲话、缓慢平移(头部做平行于相机平面的慢速移动)、快速平移(头部做平行于相机平面的快速移动)、小幅旋转(头部做左右 20 °的转动)以及中等旋转(头部做左右 35 °的转动)。视频由 SVS-Vistek GmbH 公司的 eco274 CVGE 相机拍摄,帧率为 30 帧每秒(frame per second,FPS),分辨率为 640 × 480 像素,焦距为 4.8 mm。脉搏真值采用指夹式脉搏血氧计(康泰 CMS50E)并行采集,采样率为 60 Hz。

UBFC-rPPG:该数据集使用低成本的网络摄像头(罗技 C920 HD Pro)采集,帧率为 30 FPS,分辨率为 640 × 480 像素,视频数据保存格式为未压缩的 8 位 RGB 格式。使用康泰 CMS50E 脉搏血氧计获得 PPG 波形数据。在拍摄过程中,受试者坐在摄像机前(离摄像机约 100.0 mm 远),面部清晰可见。所有实验都是在日光以及室内照明环境的室内进行。

3.2 实验细节

本文使用 Kanade-Lucas-Tomasi(KLT)算法[32-34]在连续视频帧中检测和跟踪受试者的人脸,以此作为感兴趣区域(region of interest,ROI),保持 ROI 宽高比为 3∶2,并将 ROI 调整为 192 × 128 像素。为了保持 PPG 信号的采样频率与视频帧率一致,本文将 PPG 信号降采样到 30 Hz。为了生成足够的训练样本,将视频和相应的 PPG 信号使用窗口长度为 150 帧、步长为 30 帧的滑动窗口进行分割。将视频片段和相应的 PPG 信号分别作为神经网络的输入样本和标签。

3.3 评估指标

本文使用 4 个常用的指标作为算法的评估指标,即信噪比(signal-to-noise-ratio,SNR)[10]、平均绝对误差(mean absolute error,MAE)[35]、均方根误差(root mean square error,RMSE)[36]和皮尔森相关系数 ρ[29]。其中信噪比的计算方式最早由 Haan 等人[10]提出,在 rPPG 领域被广泛采用,用于衡量网络输出的脉搏信号的最大功率谱密度与真值频率的差距,其值越大说明信号质量越好,定义如下:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{f=0.8}^{5} U(f)S^2(f)}{\sum_{f=0.8}^{5}(1-U(f))S^2(f)}\right) \tag{9}$$

式中: $S(f)$ 表示提取脉搏波形的功率谱,它是通过快速傅里叶变换(fast Fourier transform,FFT)将脉搏信号转换到频域得到的; f 为频率,单位为 Hz; $U(f)$ 为分离信号和噪声的模板:

$$U(f) = \begin{cases} 1 & f_r - \dfrac{\omega}{2} \leqslant f \leqslant f_r + \dfrac{\omega}{2} \cup 2f_r - \dfrac{\omega}{2} \leqslant f \leqslant 2f_r + \dfrac{\omega}{2} \\ 0 & \text{其他} \end{cases} \tag{10}$$

MAE 和 RMSE 是误差评价指标,其值越低说明预测效果越好,定义如下:

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N} |H_p(t) - H_l(t)| \tag{11}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(H_p(t) - H_l(t))^2} \tag{12}$$

式中:$H_p(t)$ 和 $H_l(t)$ 分别为预测脉搏信号和真值脉搏信号使用离散傅里叶变换计算得到的预测的心率值和真实心率值;N 为测试样本数。

ρ 表示测量预测心率值与真实心率值之间的线性相关关系,其值越高说明预测的心率值与真实的心率值越接近,定义如下:

$$\rho = \frac{N\sum_1^N H_p(t)H_l(t) - \sum_1^N H_p(t)\sum_1^N H_l(t)}{\sqrt{((N\sum_1^N H_p^2(t) - (\sum_1^N H_p(t))^2)(N\sum_1^N H_l^2(t) - (\sum_1^N H_l(t))^2)}} \tag{13}$$

3.4 比较算法

本文选择 9 种现有的 rPPG 提取算法进行性能比较,包括 CHROM[8]、POS[9]、PBV[20]、HR-CNN[15]、DeepPhys[16]、PhysNet[14]、MTTS-CAN[28]、EfficientPhys[29]、BigSmall[37]。其中,PBV、CHROM、POS 为传统基于模型或信号处理的方法,其他方法为基于神经网络的方法。

4 实验结果和分析

4.1 与现有方法性能比较

本节将所提网络与现有 rPPG 脉搏提取方法进行比较。将每个数据集分成训练/验证/测试集,将基于神经网络的方法在训练/验证集上进行训练,并在测试集上进行测试,在测试集上给出子类别和总体结果。PURE、UBFC-rPPG 数据集的结果分别见表 2 与表 3。其中,bpm 表示每分钟心跳次数(beats per minute)。

表 2 在 PURE 数据集上的性能对比

算法	静止				讲话				缓慢平移				快速平移			
	SNR/dB	MAE/bpm	RMSE/bpm	ρ	SNR/dB	MAE/bpm	RMSE/bpm	ρ	SNR/dB	MAE/bpm	RMSE/bpm	ρ	SNR/dB	MAE/bpm	RMSE/bpm	ρ
CHROM	7.10	2.48	4.96	0.65	3.26	4.58	10.88	0.42	3.71	3.81	6.60	0.65	1.27	8.96	12.85	0.17
PBV	6.85	2.31	33.39	0..82	3.56	3.94	7.24	0.62	3.68	3.41	6.14	0.66	0.92	9.26	12.26	0.19
POS	7.84	2.04	2.82	0.88	4.11	2.90	5.35	0.69	4.43	3.42	6.11	0.70	1.66	8.45	11.64	0.28
HR-CNN	7.48	1.78	2.27	0.83	3.52	2.57	3.98	0.57	6.71	1.38	2.33	0.72	5.00	1.48	1.87	0.64
DeepPhys	3.83	3.91	10.99	0.23	1.01	2.19	7.77	0.66	1.00	3.34	11.00	0.39	-0.91	7.71	19.54	0.22
PhysNet	10.17	1.68	2.53	0.91	8.75	0.98	1.34	0.97	9.29	0.91	1.27	0.99	9.02	0.73	1.03	0.99
MTTS-CAN	5.09	3.80	10.91	0.24	1.12	2.16	6.58	0.77	3.88	2.37	9.24	0.58	1.66	6.56	18.31	0.35
BigSmall	8.41	2.35	3.69	0.86	7.69	1.51	2.34	0.88	7.33	1.46	2.68	0.85	7.01	3.05	6.03	0.80
EfficientPhys	9.37	1.88	2.88	0.85	8.42	1.36	2.56	0.94	8.01	1.78	2.76	0.87	8.33	2.68	5.45	0.88
本文方法	12.64	0.59	1.02	0.90	11.46	0.83	0.47	0.90	11.90	0.72	0.92	0.89	11.81	0.80	1.10	0.90

算法	小幅旋转				中等旋转				总体			
	SNR/dB	MAE/bpm	RMSE/bpm	ρ	SNR/dB	MAE/bpm	RMSE/bpm	ρ	SNR/dB	MAE/bpm	RMSE/bpm	ρ
CHROM	4.81	3.29	7.59	0.44	1.03	7.97	12.93	0.14	3.47	5.25	9.84	0.35
PBV	5.85	2.23	5.07	0.69	2.12	5.27	7.76	0.40	3.74	4.51	7.64	0.49
POS	6.00	2.72	5.95	0.64	3.01	5.90	7.93	0.43	4.43	4.33	7.29	0.54
HR-CNN	5.07	2.01	2.56	0.78	3.24	2.30	3.45	0.92	5.18	1.89	2.99	0.90
DeepPhys	3.02	2.68	10.23	0.34	1.71	2.23	7.8	0.35	1.52	3.44	12.14	0.33
PhysNet	9.76	0.68	0.95	0.99	8.73	0.96	1.25	0.98	9.28	0.98	1.47	0.98
MTTS-CAN	3.66	2.67	10.13	0.35	1.85	1.82	5.87	0.51	2.85	3.30	11.17	0.42
BigSmall	8.57	1.39	4.09	0.97	6.22	3.15	5.67	0.92	7.32	2.59	5.52	0.88
EfficientPhys	8.60	1.29	2.50	0.98	8.01	2.68	5.07	0.98	8.55	1.96	2.58	0.94
本文方法	11.46	0.61	0.91	0.99	10.32	0.76	0.98	0.99	11.62	0.67	0.97	0.98

从表 2 可以看出,本文方法的 SNR 在子类别和总体结果相较于所有对比方法均有明显优势。SNR 增幅最大的子类是“缓慢平移”,与排名第 2 的算法 PhysNet 相比,增幅为 30.97%。其他类别中与次优算法相比,最低增幅为 17.41%。与 PhysNet 相比,平均指标增长 25.21%。对于 MAE 和 RMSE,所提网络在“静止”“讲话”“缓慢平移”“小幅旋转”和“中等旋转”方面优于所有对比方法,在“快速平移”方面略差于 PhysNet,但本文方法的误差在总体结果中是最小的。对于 ρ,本文方法与 PhysNet 和 EfficientPhys 结果相当,均取得了较高的分数。

从表 3 可以看出,本文方法的所有性能指标均大幅优于所有对比方法。提升幅度最大的性能指标是 RMSE,与次优算法 PhysNet 相比,下降幅度为 47.05%。对于 MAE,本文方法优于所有对比方法,比次优方法降低 15.38%。对于 SNR 和 ρ,本文方

法分别比次优方法增长 13.62% 和 0.40%。

表 3　在 UBFC-rPPG 数据集上的性能对比

算法	SNR/dB	MAE/bpm	RMSE/bpm	ρ
CHROM	0.61	9.58	20.01	0.630
PBV	0.30	11.51	22.04	0.550
POS	1.19	7.64	17.03	0.710
HR-CNN	5.45	1.86	2.66	0.960
DeepPhys	3.37	3.66	14.75	0.770
PhysNet	6.53	0.78	1.87	0.995
MTTS-CAN	3.79	3.40	14.61	0.770
BigSmall	6.03	2.19	5.02	0.880
EfficientPhys	6.41	1.77	3.24	0.960
本文方法	7.42	0.66	0.99	0.999

图 5 为对比不同方法输出脉搏信号频谱图，横坐标为时间，范围为 0 ~ 60 s；纵坐标为频率，范围为 0 ~ 7 Hz。颜色越亮代表频率越高，颜色越深代表频率越低。非主频区越干净，噪声越小。从图中可以看出主导频率（心率）与其高次谐波随时间的变化趋势。与其他方法相比较，本文方法的优点有以下 3 个方面：首先，一些方法在低频段内含有相当强度的噪声，如 CHROM、PBV、POS 和 HR-CNN，这是由于被拍摄者头部转动、身体移动或说话等情形，算法受到干扰所致，也是导致心率估计误差的主要原因。相比之下，本文方法在低频段的信号强度很低，对低频噪声的抑制体现了算法对运动伪影的去除能力。其次，传统的基于模型的方法性能下降的原因是针

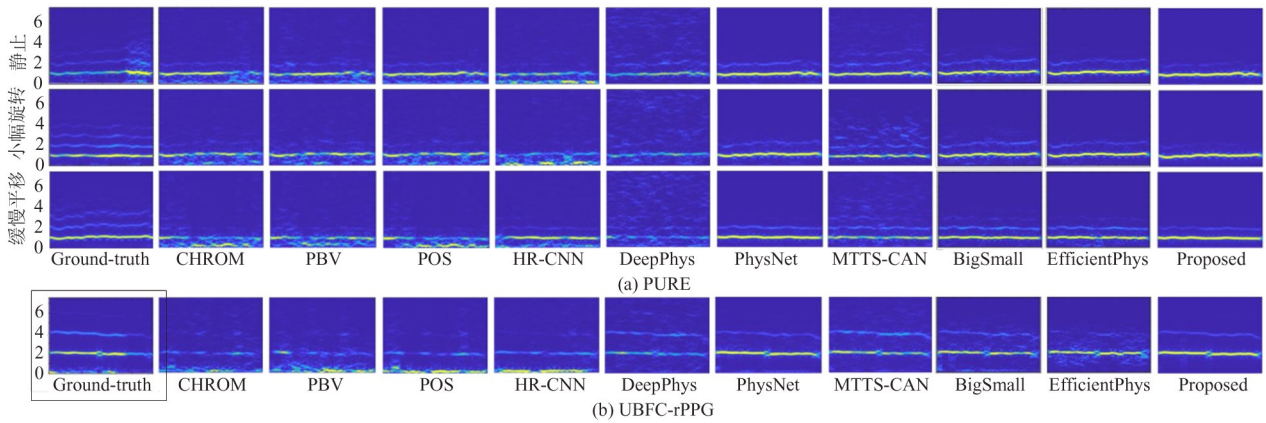

图 5　输出波形频谱图比较

对新的拍摄环境，模型的参数没有能够及时调整，对于 HR-CNN 等深度学习算法来说，则是因为没有有效处理数据的帧间关系所致，而本文方法通过 STC-JANet 较好地处理好了帧间关系。最后，一些方法存在大量高频噪声，如 HR-CNN、DeepPhys、BigSmall 和 MTTS-CAN，而本文方法在高频段内信号强度均很低。高频噪声的来源主要为相机内部成像器件。DeepPhys、BigSmall 和 MTTS-CAN 为了对脉搏随时间的变化进行建模在算法内部引入了帧间差分，这同样也放大了高频噪声。而本文方法通过对时间窗口进行不同尺度的处理，综合了全局时间和局部时间特征的优势，在保证了帧间关联的同时抑制了高频噪声。

图 6 展示了估计脉搏波形与真实波形的比较。可以看出，一些方法的脉搏波形表现为低频趋势，如 CHROM、PBV、POS。基于卷积网络的方法，如 DeepPhys、MTTS-CAN，表现出大量的高频噪声。相比之下，本文方法的波形与真实波形有更高的拟合度。

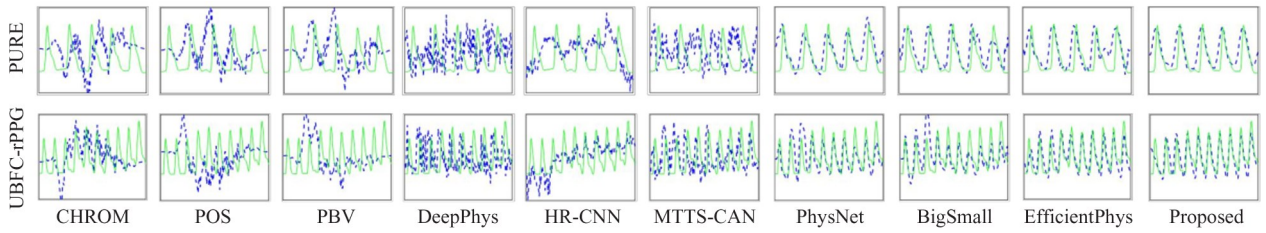

图 6　估计脉搏波形（虚线）和真实脉搏波形（实线）的比较

图7展示了估计心率与真实心率的比较。图7中2条线的重叠程度越高,表明算法的精度越好。由图可见,对比方法的心率曲线与真实心率曲线存在较大偏差,表现为低于真实心率(由运动伪影干扰)和高于真实心率(由高频噪声干扰)2种情形。与其他算法相比,本文方法的心率曲线在2个数据集上与真实心率曲线的重叠程度最高,也说明了所提算法的有效性。

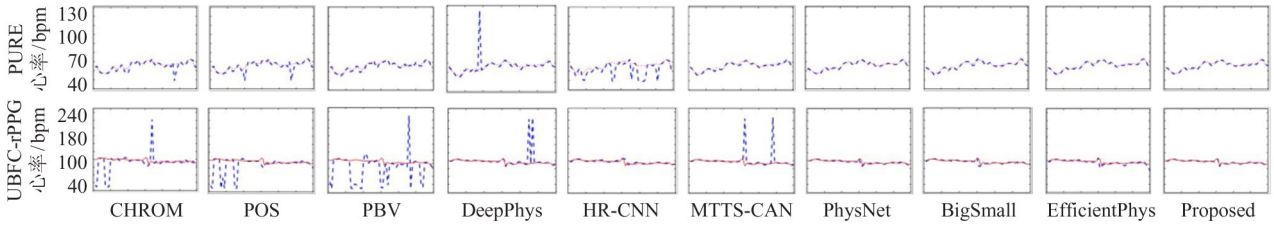

图7 估计心率(虚线)与真实心率(实线)的比较

4.2 与现有方法的复杂度比较

对本文所提对比深度学习的算法进行了模型的复杂度分析与比较,结果如表4所示。参数量代表网络模型中需要训练的参数总数;存储是代表深度学习完成后保存的模型文件大小;FLOPs(floating point operations)为浮点运算数。

表4 不同算法的模型复杂度比较

方法	输入尺寸	参数量/$\times10^6$	存储/MB	FLOPs/$\times10^9$
DeepPhys	$3\times150\times36\times36$	1.46	5.70	9.62
HR-CNN	$3\times300\times192\times168$	1.87	7.32	988.67
MTTS-CAN	$3\times150\times36\times36$	1.45	5.70	9.61
PhysNet	$3\times128\times128\times128$	0.83	3.26	130.52
BigSmall	$3\times3\times144\times144$	2.14	8.20	1.37
EfficientPhys	$3\times300\times72\times72$	2.16	8.27	35.00
本文方法	$3\times150\times192\times168$	0.37	1.45	70.58

这3个指标一同用来衡量算法/模型的复杂度与大小。经过实验获取各参数,实验结果证明本文算法的模型参数量和存储大小在7种深度学习算法中是最低的,但是FLOPs较高。这说明本文方法模型空间复杂度较低,显存和本地文件占用空间小,但是时间复杂度较高,进行深度学习训练中需要GPU较高的利用率。

4.3 关键成分的消融实验

本节对所提注意力网络模块STCJANet中的2个子模块STWNet和GTNet进行消融实验研究,以检验其对整体网络性能的影响。当STWNet与GTNet均不存在时,将对应的损失函数置0。在PURE数据集和UBFC-rPPG数据集的实验结果分别见表5与表6。

表5 在PURE数据集上对网络中的注意力模块的消融实验结果

	STWNet	GTNet	SNR/dB	MAE/bpm	RMSE/bpm	ρ
$STCJANet_{00}$	×	×	7.97	0.96	1.33	0.892
$STCJANet_{10}$	√	×	9.81	0.86	1.21	0.916
$STCJANet_{01}$	×	√	9.06	0.93	1.73	0.887
$STCJANet_{11}$	√	√	11.53	0.67	0.97	0.978

表6 在UBFC-rPPG数据集上对网络中的注意力模块的消融实验结果

	STWNet	GTNet	SNR/dB	MAE/bpm	RMSE/bpm	ρ
$STCJANet_{00}$	×	×	5.79	0.96	2.40	0.991
$STCJANet_{10}$	√	×	5.19	0.93	1.88	0.995
$STCJANet_{01}$	×	√	6.17	1.14	2.19	0.993
$STCJANet_{11}$	√	√	7.42	0.66	0.99	0.999

从实验结果可以发现，对于 PURE 数据集，当 STCJANet 不存在时（$STCJANet_{00}$），所有指标均差于存在或部分存在 STCJANet 的性能；当仅增加 STW-Net 或 GTNet 时（$STCJANet_{10}$与 $STCJANet_{01}$），所有性能指标均有所提升；当二者同时存在时（$STCJANet_{11}$），网络性能达到了最优。在 UBFC-rPPG 数据集上得到了相同的变化趋势。以上实验说明了本文所提的分离时间窗口网络模块和全局时域网络模块 2 种联合注意力机制的有效性。

随后对损失函数的权重系数进行消融实验研究，以观察其对网络整体性能的影响。进行对损失函数对应系数设定的研究，保持输出损失项 $Loss_o$ 的系数恒定为 1，改变 STCJANet 损失项的系数 α、β、θ。在 PURE 数据集和 UBFC-rPPG 数据集上的实验结果分别见表 7 和表 8。

从实验结果中可以看出，当 α、β、θ 分别为 0.1、0.2、0.3 时，网络性能均达到最优，无论是对 PURE 数据集还是 UBFC 数据集。值得注意的是，权重系数并非取值相同时网络性能达到最好，而是在逐步递增的。这可以解释为数据在网络传播的过程中，不同深浅程度的特征对最终网络性能的影响不同。对于浅层特征，其 rPPG 脉搏波动性并不明显，因此在损失函数中所占比重小；随着网络的深入，rPPG 特征逐渐显现，在损失函数中的占比也随之增加。

表 7　在 PURE 数据集上对损失函数系数的消融实验结果

	α	β	θ	SNR/dB	MAE/bpm	RMSE/bpm	ρ
Loss	0.1	0.1	0.1	8.25	0.75	1.07	0.915
	0.2	0.2	0.2	10.00	0.77	1.08	0.925
	0.3	0.3	0.3	10.81	0.73	1.02	0.930
	0.1	0.2	0.3	11.53	0.67	0.97	0.978

表 8　在 UBFC-rPPG 数据集上对损失函数权重系数的消融实验结果

	α	β	θ	SNR/dB	MAE/bpm	RMSE/bpm	ρ
Loss	0.1	0.1	0.1	6.82	0.72	1.08	0.998
	0.2	0.2	0.2	6.97	0.78	1.27	0.998
	0.3	0.3	0.3	7.09	0.77	1.37	0.998
	0.1	0.2	0.3	7.42	0.66	0.99	0.999

最后，对 3 个 STCJANet 损失项进行消融研究，以探究 $Loss_i$ 存在与否对性能的影响。保持 α、β、θ 初始为全 0，依次增加 $Loss_3$、$Loss_2$、$Loss_1$ 的对应权重系数，实验结果分别见表 9 和表 10。

表 9　在 PURE 数据集上对损失函数系数对网络影响的消融实验结果

	α	β	θ	SNR/dB	MAE/bpm	RMSE/bpm	ρ
Loss	×	×	×	9.96	0.69	1.18	0.928
	×	×	√	10.56	0.69	1.01	0.942
	×	√	√	11.59	0.75	1.06	0.926
	√	√	√	11.53	0.67	0.99	0.978

表 10　在 UBFC-rPPG 数据集上对损失函数系数对网络影响的消融实验结果

	α	β	θ	SNR/dB	MAE/bpm	RMSE/bpm	ρ
Loss	×	×	×	6.84	0.82	3.04	0.987
	×	×	√	7.06	0.70	1.06	0.999
	×	√	√	7.43	0.71	1.20	0.998
	√	√	√	7.42	0.66	0.99	0.999

实验结果表明,权重系数对整体性能产生了积极的影响。对比α、β、θ均为0时,将θ置为0.3后,所有指标均变好,说明$Loss_3$损失项对整个学习过程起到了积极作用。随后,增加β值为0.2时,2个数据集上SNR性能变好,但MAE、RMSE、ρ的性能变差,说明$Loss_2$损失项对整个学习过程起到了提高信噪比的作用,但是会略微降低精准度。最后,增加α值为0.1时,在2个数据集上的SNR降低小于1.00%,但是MAE、RMSE、ρ均有大幅提升,说明$Loss_1$损失项对整个学习过程起到了提高精度的作用,但是会略微降低信噪比。

4.4 交叉数据集测试

为了测试泛化能力,在源数据集上对模型进行训练,并在目标数据集上进行测试,并将其与在目标数据集上训练和测试的模型进行比较。结果如表11所示,其中箭头表示与在源数据集上训练和测试的结果相比的趋势。

从结果中可以看出,在不同的数据集上进行训练和测试时,所有模型的SNR和ρ都下降,MAE和RMSE都增加。然而,本文方法的SNR和ρ的振幅下降最小,MAE和RMSE的振幅增加最小,表明本文方法具有一定的泛化能力。此外,研究结果还表明,在PURE数据集上进行训练时,该网络获得了SNR和ρ的最小下降幅度,以及MAE和RMSE的最小增加幅度。其原因是,PURE数据集包含突然转头、头部移出画面、被遮挡、跟踪框跳变或消失等异常情况,本文所提出的双注意力机制,可以较好地进行自注意力的变量分配,优化部分空间特征提取,从而使网络有较好的泛化能力。

表11 交叉数据集的模型测试结果

算法	源→目标	SNR/dB	MAE/bpm	RMSE/bpm	ρ
HR-CNN	URE→UBFC	1.44(↓4.01)	6.97(↑5.11)	9.86(↑7.20)	0.730(↓0.23)
DeepPhys		0.69(↓2.68)	4.99(↑1.33)	16.35(↑1.60)	0.670(↓0.10)
PhysNet		4.14(↓2.39)	1.95(↑1.17)	6.22(↑4.35)	0.987(↓0.08)
MTTS-CAN		2.64(↓1.15)	4.55(↑1.15)	16.51(↑1.90)	0.680(↓0.09)
BigSmall		4.02(↓2.01)	3.31(↑1.12)	6.73(↑1.71)	0.750(↓0.13)
EfficientPhys		3.40(↓3.01)	2.97(↑1.20)	4.82(↑1.88)	0.850(↓0.11)
本文方法		7.21(↓0.21)	1.51(↑0.05)	1.07(↑0.08)	0.979(↓0.02)
HR-CNN	UBFC→PURE	2.74(↓2.44)	5.44(↑3.55)	10.89(↑7.90)	0.750(↓0.15)
DeepPhys		0.08(↓1.44)	6.64(↑3.20)	21.74(↑9.60)	0.220(↓0.11)
PhysNet		8.50(↓0.78)	1.10(↑0.12)	1.75(↑0.28)	0.930(↓0.05)
MTTS-CAN		0.23(↓2.62)	4.95(↑1.65)	15.37(↑4.20)	0.360(↓0.06)
BigSmall		5.01(↓2.31)	4.72(↑2.14)	10.72(↑5.20)	0.770(↓0.11)
EfficientPhys		5.53(↓3.02)	4.05(↑2.09)	8.38(↑5.80)	0.860(↓0.08)
本文方法		11.16(↓0.46)	0.74(↑0.08)	1.11(↑0.14)	0.940(↓0.04)

5 结 论

本文针对视频拍摄过程中产生的异常情况,例如突然转头、头部移出画面、被遮挡、跟踪框跳变或消失等,提出一种融合部分-整体时间窗口及双注意力机制的视频心率提取网络。提出将时间窗口进行分割处理的方法,降低异常数据对网络性能的影响。提出部分-整体时间窗口相融合的注意力学习模式,弥补分割时间窗口带来的不连续性。设计时间-通道相分离的联合注意机制,并应用于部分-整体时间窗口的处理中。基于以上网络模块,设计新的rPPG脉搏提取网络,并通过注意力模块中引出变量,定义损失函数,与网络最终输出的变量共同计算损失函数,实现了稳定、准确的rPPG波形输出。在2个公开数据集上与9个现有算法进行了性能评估。所提方法在大部分情况下均优于其他算法。实验结果证明,对输入特征在时域方向进行分割处理的方法可以有效应对视频数据异常情况。融合部分-整体时

间窗口特征及时间-通道联合注意力机制均可以有效抑制噪声干扰、进一步提升心率测量精度。但存在增加网络训练所花费的计算机资源的问题,未来工作将进一步探索提高网络的鲁棒性,在保持较好性能的同时进一步减少模型的复杂度,以期获得更加高效、准确的心率识别性能,更好地适应未来的实际部署工作。

参考文献

[1] 李思冬. 中国 12 省份社区人群心血管疾病与死亡及其常见可改变危险因素人群归因风险研究[D]. 北京:北京协和医学院,2022.

[2] Wang W, Brinker A C, Stuijk C, et al. Robust heart rate from fitness videos[J]. Institute of Physics and Engineering in Medicine, 2017,38(6):1023 – 1044.

[3] Tarassenko L, Villarroel M, Guazzi A, et al. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models[J]. Physiological Measurement, 2014,35(5):807 – 831.

[4] Aarts L A. Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit—apilot study[J]. Early Human Develop, 2013, 89(12): 943 – 948.

[5] Guo Z, Wang Z J, Shen Z. Physiological parameter monitoring of drivers based on video data and independent vector analysis[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2014: 4374 – 4378.

[6] Huang B, Chen W, Lin C, et al. A neonatal dataset and benchmark for non-contact neonatal heart rate monitoring based on spatio-temporal neural networks[J]. Engineering Applications of Artificial Intelligence,2021,106:104447.

[7] Vogels T, Gastel M, Wang W, et al. Fully-automatic camera-based pulse-oximetry during sleep[C]// Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA: IEEE, 2018:1430 – 1438.

[8] 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.

[9] Lewandowska M, Rumi′nski J, Kocejko T, et al. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity[C]// Proceedings of the Federated Conference on Computer Science and Information Systems. Szczecin, Poland: IEEE, 2011:405 – 410.

[10] De Haan G, Jeanne V. Robust pulse rate from chrominance-based rPPG[J]. IEEE Transactions on Biomedical Engineering, 2013,60(10):2878 – 2886.

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

[12] Wang W, Stuijk S, De Haan G. A novel algorithm for remote photoplethysmography: spatial subspace rotation[J]. IEEE Transactions on Biomedical Engineering, 2016, 63(9):1974 – 1984.

[13] Zhao C, Mei P, Xu S, et al. Performance evaluation of visual object detection and tracking algorithms used in remote photoplethysmography[C]// Proceedings of the IEEE International Conference on Computer Vision Workshop. Seoul, Korea: IEEE,2019:1646 – 1655.

[14] Yu Z, Li X, Zhao G. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks[EB/OL]. (2019 – 07 – 31)[2024 – 04 – 11]. https://arxiv.org/pdf/1905.02419v2.

[15] Spetlik R, Franc V, Cech J, et al. Visual heart rate estimation with convolutional neural Network[C]// Proceedings of the Brtish Machine Vision Conference. Newcastle, UK: BMVA, 2018:3 – 6.

[16] Chen W, Mcduff D. Deepphys: video-based physiological measurement using convolutional attention Networks[C]// Proceedings of the European Conference on Computer Vision. Online: ECCV, 2018:349 – 365.

[17] Liu X, Fromm J, Patel S, et al. Multi-task temporal shift attention networks for on-device contactless vitals measurement[C]// Proceedings of the Conference on Neural Information Processing Systems. Online: NeurIPS, 2020: 19400 – 19411.

[18] Hu M, Qian F, Guo D, et al. Eta-rppgNet: effective time-domain attention network for remote heart rate meas-

urement[J]. IEEE Transactions on Instrumentation and Measurement, 2021,70:1 - 12.

[19] Liu S Q,Yuen P C. A general remote photoplethysmography estimator with spatiotemporal convolutional network[C]// IEEE International Conference on Automatic Face and Gesture Recognition. Buenos Aires, Argentina: IEEE, 2020:481 - 488.

[20] Yu Z, Li X, Niu X, et al. AutoHR: a strong end-to end baseline for remote heart rate measurement with neural searching[J]. IEEE Signal Processing Letters, 2020,27: 1245 - 1249.

[21] Zhao C, Wang H, Chen H, et al. JAMSNet: a remote pulse extraction network based on joint attention and multi-scale fusion[J]. IEEE Transactions on Circuits and Systems for Video Technology,2023,33(6):2783 - 2797.

[22] Hertzman A B. Photoelectric plethysmography of the fingers and toes in man[J]. Proceedings of the Society for Experimental Biology and Medicine,1937,37(3):529 - 534.

[23] Bahdanau D,Cho K,Bengio Y. Neural machine translation by jointly learning to align and translate[EB/OL]. (2015 - 05 - 19)[2024 - 04 - 11]. https://arxiv.org/pdf/1409.0473.

[24] Zhao X,Liu J,Wu X,et al. Probabilistic spatial distribution prior based attentional keypoints matching network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022,32(3):1313 - 1327.

[25] Kong J, Deng H, Jiang M. Symmetrical enhanced fusion Network for skeleton-based action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021,31(11):4394 - 4408.

[26] Zhang X,Wang T,Jiang R,et al. Multi-attention convolutional neural network for video deblurring[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022,32(4):1986 - 1997.

[27] Liu X, Hill B L, Jiang Z, et al. EfficientPhys: enabling simple, fast and accurate camera-based vitals measurement[C]// IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023:4997 - 5006.

[28] Bobbia S, Macwan R, Benezeth Y, et al. Unsupervised skin tissue segmentation for remote photoplethysmography[J]. Pattern Recognition Letters, 2019,124:82 - 90.

[29] Shi J,Tomasi C. Good features to track[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle,USA:IEEE,1994:593 - 600.

[30] Fu F, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE,2019:3141 - 3149.

[31] Tomasi C,Kanade T. Detection and tracking of point features[C]// International Journal of Computer Vision, 1991,9(3):37 - 154.

[32] Clevert D A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units[EB/OL]. (2015 - 12 - 03)[2024 - 04 - 11]. https://arxiv.org/pdf/1511.07289v2.

[33] Huang P W, Wu B J, Wu B F. A heart rate monitoring framework for real-world drivers using remote photoplethysmography[J]. IEEE Journal of Biomedical and Health Informatics, 2021,25(5):1397 - 1408.

[34] Liu Y,Qin B,Li R,et al. Motion-robust multimodal heart rate estimation using BCG fused remotePPG with deep facial ROI tracker and pose constrained kalman filter[J]. IEEE Transactions on Instrumentation and Measurement, 2021,70(99):1 - 15.

[35] Narayanswamy G, Liu Y, Yang Y. Bigsmall: efficient multi-task learning for disparate spatial and temporal physiological measurements[C]// IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE, 2024:7914 - 7924.

[36] Hu J,Shen L,Sun G. Squeeze-and-excitation networks[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City,USA:IEEE,2018:7132 - 7141.

[37] Meng L, Zhao B, Chang B, et al. Interpretable spatio-temporal attention for video action recognition[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea: IEEE, 2019:1513 - 1522.

Video-based heart rate extraction network via partial-global time window dual attention fusion

Zhao Changchen[*,**], Cao Pengcheng[*], Huang Bin[***], Feng Yuanjing[*]

([*] College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023)

([**] School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018)

([***] Hangzhou Innovation Institute, Beihang University, Hangzhou 310051)

Abstract

While numerous remote photoplethysmography (rPPG) heart rate extraction neural networks have emerged in recent years, most of them only consider spatiotemporal continuity and neglect unpredictable anomalies in video data, leading to a decrease in heart rate measurement accuracy. This paper proposes a video-based heart rate extraction network based on the fusion of partial-global temporal window features and a time-channel joint attention mechanism. Its aim is to accurately estimate heart rate in the presence of abnormal situations in the video, and to define a hybrid loss function by combining the attention mechanism loss. The proposed network is compared with nine existing methods on PURE and UBFC-rPPG datasets to validate its performance. Experimental results show that on the PURE dataset, compared with the second-best model, signal-to-noise-ratio (SNR) of the proposed network is improved by 25.21%, mean absolute error (MAE) decreased by 31.63%, root mean square error (RMSE) decreased by 34.00%, and ρ increased by 4.25%. On the UBFC-rPPG dataset, compared to the second-best model, the SNR of the proposed network is improved by 13.62%, MAE decreased by 15.38%, RMSE decreased by 47.05%, and ρ increased by 0.40%. The experimental results demonstrate that the fusion of partial-global temporal window features and the time-channel joint attention mechanism can effectively suppress noise interference and further enhance heart rate measurement accuracy.

Key words: remote photoplethysmography, neural network, heart rate estimation, separated time window, video processing