

DOI: 10.13382/j.jemi.B2609203

# 融合掩码自编码器与对比学习的视频心率检测<sup>\*</sup>

王家雄<sup>1</sup> 李佳杰<sup>1</sup> 范伟<sup>1</sup> 宋仁成<sup>1</sup> 刘羽<sup>1</sup> 成娟<sup>1,2</sup>

(1. 合肥工业大学仪器科学与光电工程学院 合肥 230009; 2. 合肥工业大学测量理论与精密仪器安徽省重点实验室 合肥 230009)

**摘要:** 远程光电容积描记法(remote photoplethysmography, rPPG)作为一种非接触式心率检测技术,广泛应用于健康监护、情感计算和驾驶安全等领域。现有 rPPG 监督学习方法高度依赖真实生理信号标签,而参考标签获取难度大,标注成本高,制约其实际应用。针对上述问题,提出一种融合掩码自编码器(masked autoencoder, MAE)与对比学习(contrastive learning, CL)的自监督学习方法,记为 PhysCMAE,实现参考生理信号标签数量有限场景下视频心率的准确估计。首先,基于面部皮肤感兴趣区域像素值随时间变化构建多尺度时空图;其次,设计基于视觉 Transformer 的非对称编码器-解码器结构,对多尺度时空图进行掩码自编码器预训练,以学习细粒度生理特征表示;随后,在预训练后期引入对比学习模块,通过频域对比策略增强生理特征的实例可区分性;最后,添加 rPPG 预测头并对预训练后的编码器进行微调,实现从视频生理特征到高精度血容量脉搏(blood volume pulse, BVP)信号的映射及心率估计。所提方法在两个公开数据集 UBFC-RPPG 和 PURE 的集内实验中,心率平均绝对误差分别为 0.71 bpm 和 0.59 bpm;跨集实验中,心率平均绝对误差分别为 3.80 bpm 和 4.88 bpm,较基线方法其误差性能分别提升约 68.0%和 15.2%左右。在自采数据集 BSIPL-RPPG 的跨集实验中,心率平均绝对误差分别为 3.96 bpm 和 6.69 bpm,与现有监督学习方法相比达到次优结果。实验结果表明,PhysCMAE 相较现有自监督学习方法能够在生理信号标签稀缺条件下具有较高的心率检测精度与跨集泛化能力,为视频 rPPG 技术的实用化应用提供了一种有效的技术方案。

**关键词:** 远程光电容积描记法;心率检测;自监督;掩码自编码器;对比学习

**中图分类号:** TN911.7; TP751.1      **文献标识码:** A      **国家标准学科分类代码:** 310.61

## Self-supervised video-based heart rate measurement via combining masked autoencoders and contrastive learning

Wang Jiaxiong<sup>1</sup> Li Jiajie<sup>1</sup> Fan Wei<sup>1</sup> Song Rencheng<sup>1</sup> Liu Yu<sup>1</sup> Cheng Juan<sup>1,2</sup>

(1. Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China; 2. Anhui Province Key Laboratory of Measuring Theory and Precision Instrument, Hefei University of Technology, Hefei 230009, China)

**Abstract:** Remote photoplethysmography (rPPG), as a contactless heart rate measurement technique, has been widely applied in health monitoring, affective computing, and driving safety. Existing supervised rPPG methods rely heavily on ground truth physiological signal labels, yet such reference labels are difficult to acquire and costly to annotate, which limits their practical deployment. To address this issue, this paper proposes a self-supervised learning method that integrates masked autoencoder (MAE) with contrastive learning (CL), termed PhysCMAE, to enable accurate video-based heart rate measurement when only limited reference physiological signal labels are available. First, multi-scale spatiotemporal maps are constructed from the temporal variations of pixel values in facial skin regions of interest. Next, an asymmetric encoder-decoder architecture based on a Vision Transformer is designed to perform masked autoencoder pretraining on the multi-scale spatiotemporal maps, thereby learning fine-grained physiological feature representations. Then, in the later stage of pretraining, a contrastive learning module is introduced, and a frequency-domain contrastive strategy is adopted to enhance the instance discriminability of physiological features. Finally, an rPPG prediction head is added, and the pretrained encoder is finetuned to map video physiological features to high-precision blood volume pulse (BVP) signals and perform heart rate estimation. In within-dataset

experiments on two public datasets, UBFC-RPPG and PURE, the proposed method achieves mean absolute errors of 0.71 bpm and 0.59 bpm, respectively. In cross-dataset experiments, the mean absolute errors are 3.80 bpm and 4.88 bpm, representing error reductions of about 68.0% and 15.2%, respectively, compared with baseline methods. In cross-dataset experiments on the self-collected BSIPL-RPPG dataset, the mean absolute errors are 3.96 bpm and 6.69 bpm, respectively, achieving the second-best results among existing supervised learning methods. Experimental results demonstrate that, compared with existing self-supervised learning approaches, PhysCMAE can achieve higher heart rate estimation accuracy and stronger cross-dataset generalization under conditions of scarce physiological signal labels, providing an effective technical solution for the practical application of video-based rPPG technology.

**Keywords:** remote photoplethysmography; heart rate detection; self-supervised; masked autoencoders; contrastive learning

## 0 引言

心率 (heart rate, HR)、心率变异性 (heart rate Variability, HRV)、呼吸频率 (respiratory rate, RR) 等生理参数是评估个体健康状况的重要指标。由于心脏是人体最重要的器官之一,心率的估计和监测对于心血管和慢性疾病的治疗至关重要<sup>[1]</sup>。传统心率测量方法主要通过心电图、指夹式血氧仪等接触式生理监测设备。这些设备精度较高,但存在舒适性差、限制活动、难以长期连续监测等局限。近年来,基于视频的非接触式生理参数测量技术,因其非接触式、便捷、低成本的特性成为研究热点<sup>[2]</sup>。

远程光电容积描记法 (remote photoplethysmography, rPPG) 是基于视频的非接触式生理参数测量技术的核心实现手段。其通过分析视频中由于心脏搏动引发的人体面部区域的细微光度变化,间接提取心率<sup>[3]</sup>。目前已有许多基于传统信号处理与分析的方法被提出<sup>[2]</sup>。随着深度学习技术的发展,基于监督学习的 rPPG 方法在脉搏波信号质量提升方面取得了显著进展<sup>[4]</sup>。但这些方法往往依赖接触式传感器同步采集的信号作为参考标签进行训练,存在以下瓶颈:一方面,监督学习方法所需的参考标签获取难度大、标注成本高,制约其实际应用<sup>[5]</sup>;另一方面,由于脉搏传导时间 (pulse transit time, PTT) 存在个体差异,使得模型难以训练。近年来,自监督学习方法<sup>[6]</sup>引起了广泛的关注,其旨在从大量无标签样本中学习底层特征和模式,主要包括基于对比学习 (contrastive learning, CL) 的方法和基于非对比学习的方法。

对比学习作为自监督学习的重要范式<sup>[7-8]</sup>,其核心思想是通过构建正负样本对并优化特征空间的距离度量,使模型能够捕捉具有区分性的特征。在 rPPG 任务中,这类方法通过数据增强或样本间的关系假设来定义正负样本<sup>[9-11]</sup>,旨在拉近相同生理状态样本的不同增强视角的特征表示,同时推开不同生理状态样本的特征表示,从而学习对心率变化等生理信号具有鲁棒性的表征。同时,基于 CL 的方法往往通过对原始输入进行频率重采样来生成不同生理状态的负样本<sup>[12-13]</sup>,有效拓展了模型所见样本的生理状态分布。因此,模型学习到的特征具备更

强的泛化能力,能够在一定程度上减少对特定训练数据分布过拟合的风险。然而,不同视频中包含不同的噪声信号,这些噪声信号往往比生理信号更明显。只关注样本彼此间的差异,忽略对样本内部生理信息时空特征的细粒度建模,将导致模型倾向于学习样本中更为显著的非生理噪声,降低模型心率估计的性能<sup>[14]</sup>。

基于非对比学习的方法不依赖正负样本对之间的比较,而是通过设计特定的代理任务来挖掘数据内在的结构和规律,从而学习有效的特征表示。其中,Speth 等<sup>[15]</sup>开发了一种使用周期信号先验和频域滤波作为损失函数的非对比学习方法 SiNC,为 rPPG 自监督学习方法提供了一种新的思路。另一类被广泛应用的基于掩码自编码器 (masked autoencoders, MAE) 的方法则通过对输入图像的某些区域进行掩码,通过训练重建被掩码的区域,使模型学习样本的底层细粒度特征和模式<sup>[16-17]</sup>。在 rPPG 任务中的具体实现是通过减少视频<sup>[18]</sup>或时空图 (spatio-temporal map, STMap)<sup>[14]</sup>中的信息冗余和噪声,以提升 rPPG 测量的准确性。然而,这类方法主要关注输入 STMap 内部的局部关系以完成重建任务,缺乏对不同 STMap 之间关系的建模,因此在学习可区分性特征方面能力有限。此外,这些方法通常需要使用参考生理信号对预训练编码器进行微调,将学习到的生理特征映射为 rPPG 信号以进行下游心率估计。这一过程对参考生理信号标签的心率分布较为敏感,当测试数据分布与训练分布不一致时,往往难以做出准确估计。

综上所述,本文结合两种自监督学习方法的优点,提出一种全新的 rPPG 自监督学习方法 PhysCMAE,通过融合掩码自编码器 MAE 的细粒度特征学习能力与 CL 的实例可区分能力,实现参考生理信号标签有限场景下非接触式视频心率的准确检测。

本文的主要贡献如下:

1) 提出了一种融合 MAE 与 CL 的 rPPG 自监督学习方法,兼具 MAE 对多尺度时空图 (multi-scale-spatio-temporal map, MSTMap) 内部细粒度生理信息的建模能力与 CL 对跨样本判别特征的学习能力,为参考生理信号标签有限条件下的视频心率检测提供了新的技术方案。

2) 搭建基于视觉 Transformer 的非对称编码器-解码器框架进行 MAE 预训练,提取 MSTMap 被掩码区域中的

细粒度生理特征。在 CL 中使用传统 rPPG 方法生成正样本,使用频率重采样生成负样本,引导模型学习与脉搏波最为相关的特征,同时结合分阶段训练策略引导 CL 与 MAE 协同运作,提升模型在复杂场景下的鲁棒性与泛化性。

3) 在公开数据集 UBFC-RPPG、PURE 和自采数据集 BSIPL-RPPG 上进行性能评估,并与其他典型 rPPG 方法对比,结果表明 PhysCMAE 分别在 PURE 取得了集内最好结果和跨集次好结果,以及在自采 BSIPL-RPPG

取得了跨集次好结果。为参考生理信号标签数量有限场景下的非接触式视频心率检测提供了有效解决方案。

## 1 原理和方法

PhysCMAE 总体框架如图 1 所示。该模型包含 3 个模块: MSTMap 构建与数据增强、PhysCMAE 自监督预训练、rPPG 信号预测及心率估计。

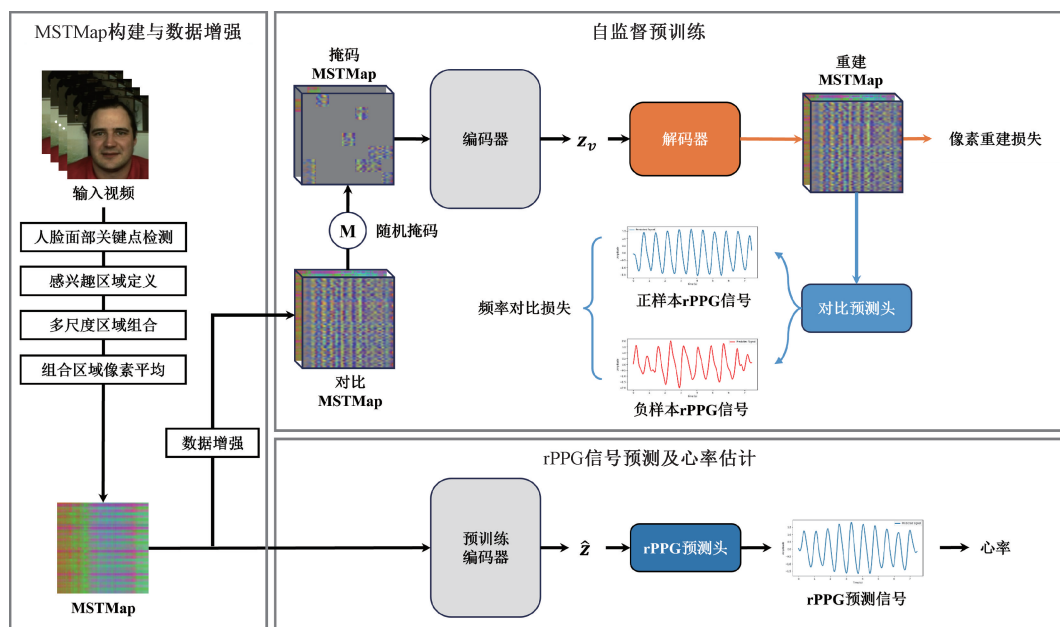


图 1 PhysCMAE 方法总框图

Fig. 1 Framework of PhysCMAE

### 1.1 MSTMap 构建与数据增强

由于 MSTMap<sup>[19]</sup> 具有排除非生理噪声的能力,因此本文方法使用 MSTMap 作为输入,通过对视频中人体面部感兴趣区域(region of interest, ROI)进行像素平均来计算。首先,利用 2D-FAN<sup>[20]</sup> 提取人脸 68 个特征点,并对缺失特征点进行插值校准。为了稳定特征点,还将移动平均滤波器应用于不同帧的面部特征点。在此基础上,定义了 6 个包含丰富生理信息的 ROI,即额头、下巴、左上脸颊、左下脸颊、右上脸颊、右下脸颊,并以  $R=2^6-1$  形式进行组合从而融合全局和局部信息。通过对整个视频中每一帧 RGB 颜色通道进行 ROI 组合及像素平均可以得到一个时间序列。对于每一个时间序列,本文方法还在 0.5~3 Hz 进行带通滤波及归一化,以减少非生理信号成分的干扰。最终对 RGB MSTMap 进行颜色空间变换,即由 RGB 空间转换为 YUV 空间,得到 YUV MSTMap 作为模型的输入。

增强是对比学习最重要的一点,相比于应用经典的

图像处理变换来轻微地改变输入,本文选择更具有生理意义的数据增强来构建正样本<sup>[13]</sup>,即通过 rPPG 传统方法对原始 RGB MSTMap 进行增强,构建传统时空图(traditional spatio-temporal map, TSTMap)作为正样本。具体来说,本文将 CHROM<sup>[21]</sup>、POS<sup>[22]</sup> 和 LGI<sup>[23]</sup> 3 种 rPPG 传统方法沿感兴趣区域 R 维应用在 MSTMap 上,并将得到的信号图沿通道维度拼接形成 TSTMap。通过使用 TSTMap 作为正样本,本文鼓励网络学习粗略的平均输入与生理上更相关的传统信号之间的相似性。对于负样本,本文方法对 YUV MSTMap 进行频率重采样来构建。选择 [0.7, 0.9] 和 [1.1, 1.3] 范围内的随机缩放因子对 YUV MSTMap 进行频率重采样,计算出频率增强时空图(frequency resample spatio-temporal map, FRSTMap)。其中, [0.7, 0.9] 对应下采样,模拟如深度睡眠休息等较低心率场景; [1.1, 1.3] 对应上采样,模拟如运动健身等较高心率场景。该范围使 FRSTMap 包含的心率仍位于正常人的合理心率范围,既避免了偏离生理意义的极端频率,又为对比学习提供了具备

一定差异的负样本。MSTMap、TSTMap 以及 FRSTMap 的构建过程如图 2 所示。

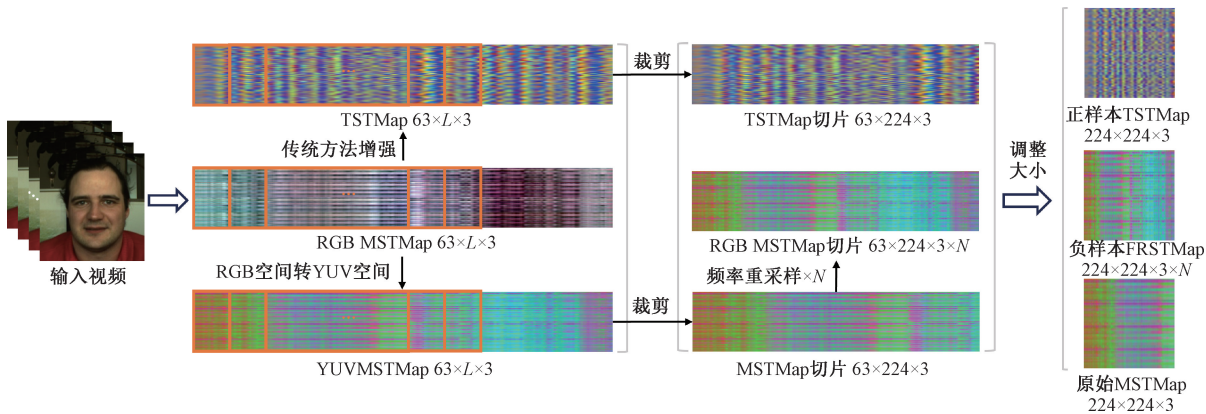


图 2 MSTMap 构建与数据增强

Fig. 2 MSTMap generation and data augmentation

## 1.2 PhysCMAE 自监督预训练

PhysCMAE 预训练流程如图 3 所示。MAE 重建需要使用掩码后的图像作为输入<sup>[16]</sup>, 本文选择随机掩码策略。掩码的实现策略很简单: 首先将输入 MSTMap 划分为不重叠的图像块, 生成一个长度为  $N$  的图像块序列  $\{x_i\}_{i=1}^N$ , 然后对图像块序列进行随机排列, 最后按比例

保留图像块序列中的部分图像块, 并删除剩余图像块。最终编码器的输入为保留的可见图像块  $X_v = \{x_v^a, x_v^p, x_v^{n_1}, x_v^{n_2}, \dots, x_v^{n_k}\}$ , 其中  $x_v^a, x_v^p$  分别代表 MSTMap 和 TSTMap 的可见图像块输入序列,  $x_v^{n_1}, x_v^{n_2}, \dots, x_v^{n_k}$  代表 FRSTMaps 的可见图像块输入序列。

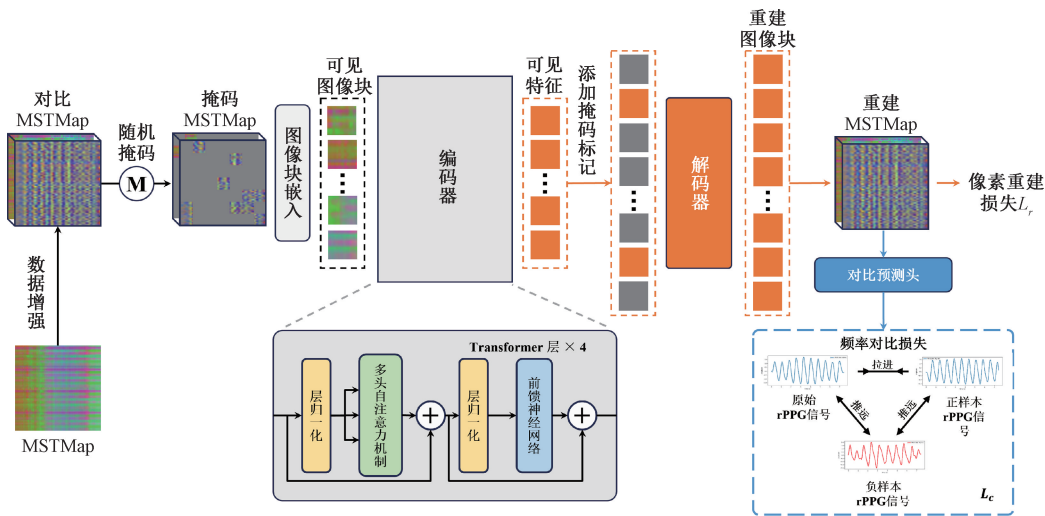


图 3 PhysCMAE 预训练

Fig. 3 PhysCMAE pre-training

### 1) 编码器

编码器采用视觉 Transformer (vision transformer, ViT) 架构。编码器  $\mathcal{F}$  首先通过线性投影将各种时空图的可见图像块  $x_v$  映射到 embedding 层, 并添加位置编码  $p_v$ , 随后将融合后的 embedding token 输入进一系列 transformer 层, 得到可见图像块的特征  $Z_v = \{z_v^a, z_v^p, z_v^{n_1}, z_v^{n_2}, \dots, z_v^{n_k}\}$ 。其中,  $z_v^a, z_v^p$  分别代表编码器从表 MSTMap、TSTMap 的可见图像块中提取的特征,  $z_v^{n_1}, z_v^{n_2}, \dots, z_v^{n_k}$  代表编码器从

FRSTMaps 的可见图像块中提取的特征, 如式 (1) 所示。

$$z_v = \mathcal{F}(x_v + p_v) \quad (1)$$

根据过于复杂的模型将对自监督学习产生负面影响的这一观察<sup>[15]</sup>, 本文没有使用 MAE 中常使用的 ViT 基础模型作为编码器, 而是通过减少 ViT 中 Transformer 的层数来搭建编码器。通过限制模型的参数量使模型在有限的学习能力下优先学习不同输入样本中存在的底层通用生理特征。因此, 在所提出的方法中选择将 ViT 缩减

为4个Transformer层作为编码器,维度为768。通过预训练像素重建任务,编码器 $\mathcal{F}$ 学习到MSTMap中生理信号的底层通用生理特征,同时也将作为特征提取器,应用于后续的心率估计。

## 2) 解码器

编码器的输出只是可见图像块的特征,要重建出原始图像,需要一个解码器将各种时空可见图像块的特征 $z_v$ 和各自对应的掩码token $z_m$ 映射到与编码器和原始图像相同的特征空间。具体来说,解码器 $\mathcal{G}$ 通过学习可见图像块间的相似性,重建出被掩码图像块的像素,其同时接收可见图像块的特征和掩码token。其中,掩码token $z_m$ 是一些具有位置嵌入的可学习向量,经过解码器后,它们被重建为时空图中被掩码的图像块, $\mathbf{Y} = \{y^a, y^p, y^{n_1}, y^{n_2}, \dots, y^{n_k}\}$ 为最终重建时空图,其中 $y^a, y^p$ 分别代表重建MSTMap、TSTMap,  $y^{n_1}, y^{n_2}, \dots, y^{n_k}$ 代表编重建FRSTMaps。解码器同样为一系列堆叠的Transformer层,为了保持MAE编码器-解码器的非对称结构以及解码器更轻量的特点,本文在所提出的方法中对解码器也进行了相应的层数缩减,默认的解码器具有2个Transformer层,维度为128,重建时空图的公式如式(2)所示。

$$y = \mathcal{G}(z_v, z_m) \quad (2)$$

## 3) 对比预测头

为了使编码器提取到的特征具有实例可区分性,本文在解码器后增加了一个用于对比学习的对比预测头 $\mathcal{H}_c$ ,其结构为一层线性层。MSTMap、TSTMap以及FRSTMaps的可见图像块经过编码器-解码器后得到各自对应的重建时空图 $\mathbf{Y} = \{y^a, y^p, y^{n_1}, y^{n_2}, \dots, y^{n_k}\}$ 。首先,先将重建时空图按照ROI维度进行平均得到三通道信号 $s'$ ,随后通过对比预测头进行通道融合将其映射为rPPG信号,得到输出rPPG信号 $s^a$ 、正样本rPPG信号 $s^p$ 以及负样本rPPG信号 $\mathbf{S}^n = \{s^{n_1}, s^{n_2}, \dots, s^{n_k}\}$ ,并用于计算后续的对损失。对比预测头的输出如式(3)所示。

$$s = \mathcal{H}_c(s') \quad (3)$$

## 4) 训练损失

对于MAE重建损失,本文在重建任务中使用原始像素作为重建目标,采用均方误差(mean squared error, MSE)作为损失函数,仅计算掩码图像块的解码器重建像素与原始像素之间的损失,公式如式(4)所示。

$$L_r = \frac{1}{N_m} \sum (y - y_{gt})^2 \quad (4)$$

式中: $L_r$ 为MAE像素重建损失; $N_m$ 为时空图中被掩码图像块的数量; $y$ 和 $y_{gt}$ 分别为解码器重建时空图的掩码图像块和原始输入时空图中对应的图像块。

对于对比损失,本文在对比预测头输出的正负rPPG信号之间进行对比学习,包括输出rPPG信号 $s^a$ 、正样本

rPPG信号 $s^p$ 以及若干负样本rPPG信号 $\mathbf{S}^n = \{s^{n_1}, s^{n_2}, \dots, s^{n_k}\}$ 。由于心率检测主要与rPPG信号主频相关,因此本文主要关注输出rPPG信号在频域之间的差异。具体而言,首先计算输出rPPG信号的归一化功率谱密度(power spectral density, PSD),并将0.5~3 Hz频带外的不相关值设置为0,PSD描述了信号中不同频率成分的功率,直观展现出不同rPPG信号在频域间的差异。为了最大化输出rPPG信号和正样本rPPG信号之间的相似性并最小化输出rPPG信号和负样本rPPG信号之间的相似性,本文选择余弦相似性作为PSD之间的距离度量,并使用InfoNCE损失函数作为频率对比损失函数, $L_c$ 的定义如式(5)所示。

$$L_c = \log\left(\frac{\exp(d(f, f^p)/\tau)}{\sum_{i=1}^k (\exp(d(f, f_i^n)/\tau) + \exp(d(f^p, s_i^n)/\tau))} + 1\right) \quad (5)$$

式中: $f$ 表示信号的PSD计算结果; $d(\cdot, \cdot)$ 表示余弦相似度量; $\tau$ 表示温度系数。通过最小化 $L_c$ ,网络会缩小 $s^a$ 和正样本 $s^p$ 在频域之间的距离,并增大 $s^a$ 和所有负样本 $\mathbf{S}^n = \{s^{n_1}, s^{n_2}, \dots, s^{n_k}\}$ 在频域之间的距离。

整体的预训练损失定义为:

$$L_{pretrain} = \begin{cases} L_r & epoch < e_s \\ L_r + L_c & \text{其他} \end{cases} \quad (6)$$

其中,为避免CL与MAE训练目标冲突,本文方法在整个预训练的前期只进行MAE像素重建任务,在预训练的后期才引入CL。引入CL后的总体学习目标是重建损失 $L_r$ 和对比损失 $L_c$ 相加的组合, $e_s$ 为开始引入对比学习的训练轮次。

## 1.3 rPPG信号预测及心率估计

rPPG信号的预测通过在预训练的编码器后添加一个rPPG预测头实现,其结构为一层线性层。该阶段对预训练的编码器进行微调,并对rPPG预测头进行训练,从而将编码器预训练学习的生理特征映射为rPPG信号。值得注意的是,该阶段的输入并非为掩码MSTMap的图像块而是完整图像块,通过计算预测rPPG信号与真实BVP信号之间的负皮尔逊相关损失,实现从预训练编码器提取的特征到rPPG信号的映射,负皮尔逊相关损失计算公式如式(7)所示。

$$L_{finetune} = 1 - \text{PearC}(S_{pr}, S_{gt}) \quad (7)$$

式中: $S_{pr}$ 表示rPPG预测头输出的rPPG信号; $S_{gt}$ 表示参考真值信号。

对于后续的心率估计,本文将Welch提出的快速傅里叶变换算法应用于过滤后的rPPG信号计算出功率谱密度,选择峰值信噪比(peak signal-to-noise ratio, PSNR)

最大的对应频率为所测心率的主频,最终估计的心率值为  $60 \times f_{\max}$  bpm (beat per minute)。

## 2 数据集与评价指标

本文通过实验对提出的方法进行评估,以说明其有效性。实验共包含 3 个数据集,分别为两个公开数据集 UBFC-RPPG<sup>[24]</sup>、PURE<sup>[25]</sup>,以及一个自采数据集 BSIPL-RPPG<sup>[26]</sup>。这 3 个数据集均包含人脸视频数据与参考 PPG 信号。

### 2.1 数据集

1) UBFC-RPPG 数据集:包含 42 名受试者于真实环境下的面部视频,视频中受试者被要求玩一个与时间相关的游戏以保持心率变化,每个视频时长约 1 min。视频由 Logitech C920 HD pro 摄像头采集,帧率为 30 fps (frames per second, fps),分辨率为  $640 \times 480$  pixels。参考信号为视频拍摄过程中由 ContecMedical CMS50E 同步采集的 PPG 信号,采样率为 30 Hz。该数据集能够反映心率随时间变化场景下的算法性能。

2) PURE 数据集:包含 10 名受试者在 6 种不同设置(静止、说话、慢平移、快平移、小旋转、中等旋转)下共 60 段视频,每个视频时长约 1 min,帧率为 30 fps,分辨率为  $640 \times 480$  pixels。参考信号为视频拍摄过程中由 pulox CMS50E 同步采集的 PPG 信号,采样率为 60 Hz。本文方法通过三次样条插值将视频和相应的 PPG 信号归一化到 30 Hz。该数据集能够反映一定头部运动和随时间的光照变化干扰下的算法性能。

3) BSIPL 数据集:为本实验室自采的一个室内 rPPG 数据集,包含年龄在 18~25 岁的 37 名健康学生受试者,其中男性 24 名,女性 13 名。受试者被要求正坐在 Logitech C920 HD pro 摄像机前约 1 m 的位置,参考信号为视频拍摄过程中由 Contec-Medical CMS50E 同步采集的 PPG 信号。每个受试者拍摄一个时频,总共拍摄了 37 个视频,其中每个视频和对应的 PPG 信号时长约 4.5 min。采集过程中自然光由窗户射入造成面部的不均匀光照变化。采集的前 2 min 受试者被要求静止,剩余的 2.5 min 受试者则根据研究人员发出的指令做出相应的头部动作。视频的帧率为 30 fps,分辨率为  $640 \times 480$  pixels,PPG 信号的采样率为 30 Hz。该数据集能够反映一定剧烈头部运动和面部不均匀光照变化干扰下的算法性能。

### 2.2 评价指标

本文使用 3 个常用评价指标来评估心率估计的性能,包括平均绝对误差(mean absolute error, MeanAE),均方根误差(root mean square error, RMSE)和皮尔逊相关系

数 (pearsons correlation coefficient, R),相关定义如式(8)~(10)所示:

$$MeanAE = \frac{1}{N} \sum_{n=1}^N |HR_{\text{predict}}^{(n)} - HR_{\text{label}}^{(n)}| \quad (8)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^N (HR_{\text{predict}}^{(n)} - HR_{\text{label}}^{(n)})^2} \quad (9)$$

$$R = \frac{\sum_{n=1}^N (HR_{\text{label}}^{(n)} - \bar{X})(HR_{\text{predict}}^{(n)} - \bar{Y})}{\sqrt{\sum_{n=1}^N (HR_{\text{label}}^{(n)} - \bar{X})^2} \sqrt{\sum_{n=1}^N (HR_{\text{predict}}^{(n)} - \bar{Y})^2}} \quad (10)$$

式中:  $HR_{\text{predict}}^{(n)}$  表示由第  $n$  个时间窗口估计的 rPPG 信号计算的心率;  $HR_{\text{label}}^{(n)}$  表示由对应时间窗口的参考 PPG 信号计算的心率;  $\bar{X}$  是  $HR_{\text{label}}^{(n)}$  的均值;  $\bar{Y}$  是  $HR_{\text{predict}}^{(n)}$  的均值;  $N$  是时间窗口总数,即测试样本数。

## 3 实验与结果

### 3.1 实验设置

为了在所有实验中与其他方法进行公平的比较,本文使用了以前工作中相同的验证协议。1) 数据集内实验:对于 UBFC-RPPG 数据集使用与文献[27]中相同的协议进行训练-测试数据集的划分和评估。对于 PURE 数据集使用与文献[27-28]中相同的实验方案来划分训练集和测试集;2) 跨数据集实验:以 PURE 数据集作为训练集来测试 UBFC-RPPG 数据集、以 UBFC-RPPG 数据集作为训练集来测试 PURE 数据集、分别以 UBFC-RPPG 数据集和 PURE 数据集作为训练集来测试 BSIPL-RPPG 数据集。对于每个训练数据集,本文使用其全部数据进行预训练和微调;3) 消融实验:本文对不同正样本类型、分阶段训练策略进行消融实验,探究它们对本文方法的影响。此外,本文还进行了实验来模拟 PhysCMAE 在参考生理信号标签极度缺失情况下的心率估计性能。

由于 rPPG-MAE 使用的编码器是具有 12 层 Transformer 层的 ViT-base 模型,而本文将编码器缩减为 4 层。为了探究层数减少的影响,本文对 rPPG-MAE 的编码器也进行了层数缩减,记层数缩减后的模型为 MAE (Baseline),并进行了相应的对比实验。同时,为了满足 ViT 的输入要求,本文将完整 MSTMap 划分为若干个 MSTMap 子片段,每个子片段的长度设置为 224 对应 224 帧的时间窗口。同时为了提高模型训练的数据量,本文对训练数据进行了重叠划分,而对于测试数据,为了与其他 rPPG 方法进行公平对比,不进行重叠划分。训练数据的划分规则具体如下:对于 UBFC-RPPG 数据集和 PURE

数据集,本文以 5 帧的步长,将其划分为若干个有重叠的 MSTMap 子片段;对于 BSIPL-RPPG 数据集,本文以 10 帧的步长,将其划分为若干个有重叠的 MSTMap 子片段。

1) 预训练

将训练集输入 ViT 编码器进行训练,训练周期为 100 个 epochs,使用 Adam W 优化器,默认 beta 为 0.9 和 0.95,权重衰减为 0.05,余弦衰减的学习率调度,预热 40 个 epochs,基础学习率为 0.001。本文方法使用线性学习率缩放规则  $lr=baselr\times batchsize/256$ ,批次大小为 64。此外预训练过程中,mask 比例为 80%,负样本数量  $N$  默认设置为 3,分阶段训练策略中对比学习引入的轮次  $e$ ,默

认设置为 80。  
2) 微调  
使用参考生理信号对 ViT 编码器和 rPPG 预测器的权重进行微调,微调训练周期为 20 个 epochs。使用 Adam W 优化器,默认 beta 为 0.9 和 0.999 9,层衰减为 0.75,权重衰减为 0.05,学习率调度为余弦衰减,预热 5 个 epochs,基础学习率为 0.001。学习率缩放规则和批次大小与预训练阶段相同。

3.2 数据集内实验结果

本文方法 PhysCMAE 在 UBFC-RPPG 和 PURE 两个公开数据集上的集内实验结果如表 1 所示。

表 1 数据集内心率对比实验结果  
Table 1 Intra-dataset comparison results of HR

| 方法类别    | 方法                                   | UBFC-RPPG                   |                           |             | PURE                        |                           |             |
|---------|--------------------------------------|-----------------------------|---------------------------|-------------|-----------------------------|---------------------------|-------------|
|         |                                      | MeanAE( $\text{min}^{-1}$ ) | RMSE( $\text{min}^{-1}$ ) | R           | MeanAE( $\text{min}^{-1}$ ) | RMSE( $\text{min}^{-1}$ ) | R           |
| 传统方法    | GREEN                                | 12.85                       | 22.99                     | 0.47        | 5.36                        | 11.11                     | 0.69        |
|         | CHROM <sup>[21]</sup>                | 2.80                        | 7.64                      | <b>0.80</b> | 10.28                       | 21.29                     | 0.24        |
|         | POS <sup>[22]</sup>                  | <b>2.39</b>                 | <b>8.15</b>               | 0.78        | 4.47                        | 13.30                     | 0.64        |
|         | OMIT <sup>[29]</sup>                 | 10.30                       | 22.39                     | 0.35        | <b>2.34</b>                 | <b>7.05</b>               | <b>0.88</b> |
| 监督学习方法  | DeepPhys <sup>[30]</sup>             | 3.64                        | 12.36                     | 0.69        | 1.20                        | 4.83                      | 0.94        |
|         | PhysNet <sup>[31]</sup>              | 0.67                        | 2.49                      | <u>0.98</u> | 0.81                        | 2.92                      | 0.98        |
|         | PulseGAN <sup>[26]</sup>             | 1.19                        | 2.10                      | <u>0.98</u> | 2.28                        | 4.29                      | <b>0.99</b> |
|         | Dual-GAN <sup>[27]</sup>             | <u>0.44</u>                 | <b>0.67</b>               | <b>0.99</b> | 0.82                        | 1.31                      | <b>0.99</b> |
|         | PhysFormer <sup>[32]</sup>           | 1.09                        | 4.51                      | 0.92        | 1.68                        | 6.85                      | 0.89        |
|         | EfficientPhys <sup>[33]</sup>        | 1.27                        | 7.50                      | 0.80        | 1.27                        | 5.68                      | 0.92        |
|         | STCNet <sup>[34]</sup>               | 0.92                        | 1.52                      | <b>0.99</b> | <b>0.27</b>                 | <b>0.44</b>               | <b>0.99</b> |
|         | Contrast-Phys+(100%) <sup>[11]</sup> | 0.21                        | <u>0.80</u>               | <b>0.99</b> | <u>0.48</u>                 | <u>0.98</u>               | <b>0.99</b> |
|         | Gidon2021 <sup>[9]</sup>             | 1.85                        | 4.28                      | 0.93        | 2.30                        | 2.90                      | <u>0.99</u> |
|         | Contrast-Phys+(0%) <sup>[11]</sup>   | 0.64                        | 1.00                      | <b>0.99</b> | 1.00                        | 1.40                      | <u>0.99</u> |
| 自监督学习方法 | Yue2023 <sup>[12]</sup>              | <b>0.58</b>                 | <b>0.94</b>               | <b>0.99</b> | 1.23                        | 2.01                      | <u>0.99</u> |
|         | SiNC <sup>[15]</sup>                 | <u>0.59</u>                 | <u>1.83</u>               | 0.99        | <u>0.61</u>                 | <u>1.84</u>               | <u>0.99</u> |
|         | Zhang et al. <sup>[35]</sup>         | 1.88                        | 2.54                      | <u>0.98</u> | 1.70                        | 2.38                      | <u>0.99</u> |
|         | rPPG-MAE <sup>[14]</sup>             | 0.75                        | 1.58                      | <b>0.99</b> | 1.41                        | 5.14                      | 0.93        |
|         | MAE (Baseline) <sup>[16]</sup>       | 0.74                        | 1.57                      | <b>0.99</b> | 1.52                        | 5.45                      | 0.93        |
|         | PhysCMAE                             | 0.71                        | 1.54                      | <b>0.99</b> | <b>0.59</b>                 | <b>1.35</b>               | <b>1.00</b> |

对于 UBFC-RPPG 数据集,PhysCMAE 的 MeanAE 为 0.71 bpm,与 rPPG-MAE 和 MAE (Baseline) 相比均进一步降低了误差,但降低幅度相对有限。结果表明,在训练集与测试集分布较为一致、采集环境相对稳定、运动噪声较弱的条件下,MAE 本身已经能够较充分学习到用于心率估计的生理特征,因此引入对比学习后带来的额外收益并不显著。需要指出的是,PhysCMAE 在 UBFC-RPPG 数据集上优于大多数方法,但与部分监督方法和自监督方法相比并非绝对最优。

对于 PURE 数据集,与 UBFC-RPPG 数据集相比,其包含更明显的头部运动干扰,且不同动作条件下的噪声模式差异较大,因此更能检验模型对噪声和复杂场景的鲁棒性。在该数据集上,PhysCMAE 的 MeanAE 为 0.59 bpm,较 rPPG-MAE 和 MAE (Baseline) 均显著下降且优于

所有自监督方法与大部分监督方法,说明本文方法在复杂场景下能够更有效地抑制非生理噪声对特征学习的干扰。该结果表明,PhysCMAE 通过重建任务学习 MSTMap 内部细粒度时空特征,同时借助 TSTMap 与 FRSTMap,进一步约束模型关注与脉搏波更相关的频率特征,而非是头部运动、光照变化或个体差异等更为显著的噪声特征。尤其是在 PURE 这类包含多种运动状态的数据中,PhysCMAE 能够使模型在保持生理信息细节的同时,提高对不同生理状态的区分能力,因此比仅依赖重建目标的 MAE 更有优势。

此外,PURE 数据集心率分布范围不均匀,主要心率范围为 40~80 bpm 并包含部分 100~140 bpm 的高心率数据,导致训练数据集和测试数据集的心率分布出现了一定程度的偏差。PhysCMAE 通过对 MSTMap 进行频率

增强,进而平衡心率分布偏差,使模型在微调数据与测试数据心率分布出现偏差的情况下,也能进行正确的心率估计。

3.3 跨数据集实验

为了评估方法的可推广性,本文在 PURE、UBFC-RPPG 和 BSIPL-RPPG 之间进行了跨数据集测试。

首先,以 PURE 作为训练集来测试 UBFC-RPPG,实验结果如表 2 所示。PhysCMAE 的 MeanAE 为 3.80 bpm, 优于 rPPG-MAE 的 10.55 bpm 和 MAE (Baseline) 的 4.48 bpm。该结果表明,当 PhysCMAE 在包含较多运动扰动的 PURE 上训练后,模型能够较好地将所学特征迁移到噪声较低的 UBFC-RPPG 上。说明 PhysCMAE 在预训练阶段通过 TSTMap 和 FRSTMap 的约束学习更加稳定的生理信息。同时说明 CL 在 PhysCMAE 中发挥的作用并不是简单的增强样本差异,而是帮助模型在复杂训练数据中学习更具可迁移性的生理特征,从而在实验环境更稳定的测试集中仍保持较好性能。

表 2 在 PURE 上训练的 UBFC-RPPG  
跨数据集心率对比实验结果

| Table 2 Cross-dataset comparison results of HR while training on PURE and testing on UBFC-RPPG |                                |                              |             |
|------------------------------------------------------------------------------------------------|--------------------------------|------------------------------|-------------|
| 方法                                                                                             | MeanAE<br>(min <sup>-1</sup> ) | RMSE<br>(min <sup>-1</sup> ) | R           |
| DeepPhys <sup>[30]</sup>                                                                       | <b>1.60</b>                    | <u>7.59</u>                  | <u>0.91</u> |
| PhysNet <sup>[31]</sup>                                                                        | 3.99                           | 12.42                        | 0.78        |
| PhysFormer <sup>[32]</sup>                                                                     | 1.96                           | <b>6.73</b>                  | <b>0.93</b> |
| EfficientPhys <sup>[33]</sup>                                                                  | 1.89                           | 8.18                         | 0.90        |
| Contrast-Phys <sup>[10]</sup>                                                                  | 10.22                          | —                            | 0.45        |
| SiNC <sup>[15]</sup>                                                                           | 6.64                           | —                            | 0.59        |
| rPPG-MAE <sup>[14]</sup>                                                                       | 10.55                          | 19.92                        | 0.57        |
| MAE (Baseline) <sup>[16]</sup>                                                                 | 4.48                           | 11.88                        | 0.82        |
| PhysCMAE                                                                                       | 3.80                           | 10.05                        | 0.86        |

其次,以 UBFC-RPPG 作为训练集来测试 PURE,实验结果如表 3 所示。PhysCMAE 的 MeanAE 为 4.88 bpm,显著优于 rPPG-MAE 的 15.01 bpm 和 MAE (Baseline) 的 15.24 bpm。PURE 数据集与 UBFC-RPPG 数据集相比包含明显头部运动和更复杂噪声模式以及显著的数据分布差异。模型在相对干净、噪声较少的 UBFC-RPPG 上训练,却需要在包含明显头部运动和更复杂噪声模式的 PURE 上测试,这对模型应对噪声干扰的能力与泛化能力提出了较高要求。对于仅依赖重建目标的 MAE 而言,训练阶段学到的特征更容易受到干扰,因此在测试集噪声模式变化后性能会明显下降。而 PhysCMAE 通过在预训练后期引入对比学习,使模型不仅学习到 MSTMap 内部的细粒度生理特征,还进一步学习了跨样本的频域判别信息,因此在面对测试集分布变化时具有更强的鲁棒性。

表 3 在 UBFC-RPPG 上训练的 PURE  
跨数据集心率对比实验结果

| Table 3 Cross-dataset comparison results of HR while training on UBFC-RPPG and testing on PURE |                                |                              |             |
|------------------------------------------------------------------------------------------------|--------------------------------|------------------------------|-------------|
| 方法                                                                                             | MeanAE<br>(min <sup>-1</sup> ) | RMSE<br>(min <sup>-1</sup> ) | R           |
| DeepPhys <sup>[30]</sup>                                                                       | 7.08                           | 18.50                        | 0.66        |
| PhysNet <sup>[31]</sup>                                                                        | 11.28                          | 22.63                        | 0.57        |
| PhysFormer <sup>[32]</sup>                                                                     | 12.96                          | 23.49                        | 0.39        |
| EfficientPhys <sup>[33]</sup>                                                                  | 5.79                           | 6.64                         | 0.73        |
| Contrast-Phys <sup>[10]</sup>                                                                  | 19.61                          | —                            | 0.33        |
| SiNC <sup>[15]</sup>                                                                           | <b>4.02</b>                    | —                            | <b>0.86</b> |
| rPPG-MAE <sup>[14]</sup>                                                                       | 15.01                          | 27.03                        | 0.44        |
| MAE (Baseline) <sup>[16]</sup>                                                                 | 15.24                          | 26.80                        | 0.46        |
| PhysCMAE                                                                                       | <u>4.88</u>                    | <u>15.71</u>                 | <u>0.79</u> |

图 4 是以 UBFC-RPPG 为训练集在 PURE 上测试所得到的 Bland-Altman 图,对应表 3 中本文所提出的 PhysCMAE 方法,以及与 MAE (Baseline)、rPPG-MAE 和 PhysFormer 方法的对比结果。在 Bland-Altman 图中,两条虚线之间的间隔反映了预测值与真实值之间误差的标准差大小,而实线的纵坐标位置则体现了预测值相对于真实值的总体偏移程度。结果表明,PhysCMAE 方法在跨数据集测试中表现出更高的可靠性,其预测误差均值最小,且 95%一致性界限更集中。说明该方法预测结果与参考值之间的一致性更加集中,预测结果与参考值之间具有更小的偏移和更稳定的误差分布,也意味着 PhysCMAE 不仅在平均误差指标上更优,而且在跨数据集场景下具有更好的估计稳定性和可靠性。

表 4 BSIPL-RPPG 跨数据集心率对比实验结果

| Table 4 Cross-dataset comparison results of HR while testing on BSIPL-RPPG |      |                                |                              |             |
|----------------------------------------------------------------------------|------|--------------------------------|------------------------------|-------------|
| 方法                                                                         | 训练集  | 测试集<br>BSIPL-RPPG              |                              |             |
|                                                                            |      | MeanAE<br>(min <sup>-1</sup> ) | RMSE<br>(min <sup>-1</sup> ) | R           |
| DeepPhys <sup>[30]</sup>                                                   | UBFC | 10.25                          | 18.27                        | 0.24        |
|                                                                            | PURE | 12.11                          | 20.20                        | 0.16        |
| PhysNet <sup>[31]</sup>                                                    | UBFC | <b>3.05</b>                    | <b>7.42</b>                  | <b>0.83</b> |
|                                                                            | PURE | 9.75                           | 16.02                        | 0.34        |
| PhysFormer <sup>[32]</sup>                                                 | UBFC | 5.09                           | 10.87                        | 0.59        |
|                                                                            | PURE | <b>3.78</b>                    | <b>8.86</b>                  | <b>0.76</b> |
| EfficientPhys <sup>[33]</sup>                                              | UBFC | 9.11                           | 17.37                        | 0.26        |
|                                                                            | PURE | 8.82                           | 17.05                        | 0.29        |
| rPPG-MAE <sup>[14]</sup>                                                   | UBFC | 4.70                           | 11.03                        | 0.70        |
|                                                                            | PURE | 8.85                           | 15.75                        | 0.29        |
| MAE (Baseline) <sup>[16]</sup>                                             | UBFC | 4.28                           | 10.03                        | 0.74        |
|                                                                            | PURE | 7.17                           | 14.66                        | 0.45        |
| PhysCMAE                                                                   | UBFC | <u>3.96</u>                    | <u>9.55</u>                  | <u>0.76</u> |
|                                                                            | PURE | <u>6.69</u>                    | <u>14.51</u>                 | <u>0.51</u> |

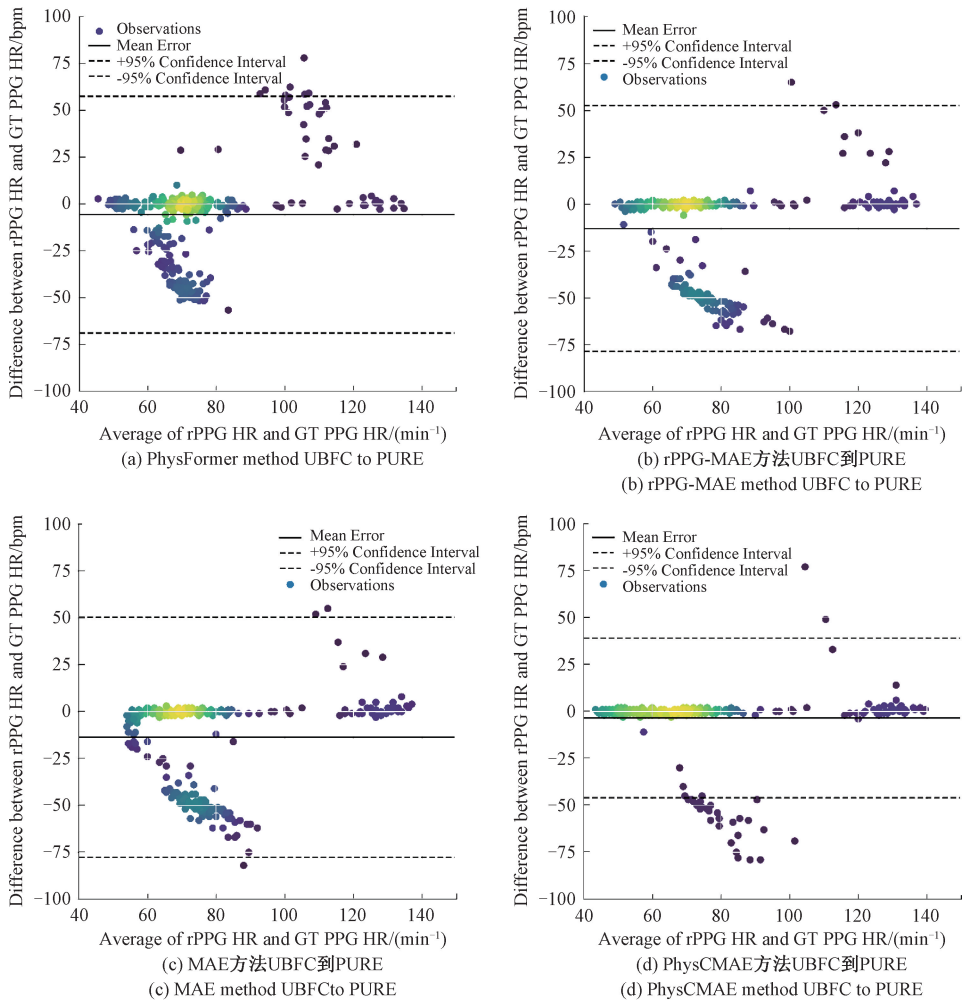


图 4 在 UBFC-RPPG 上训练的 PURE 跨数据集实验 Bland-Altman 图

Fig. 4 The Bland-Altman figures of PURE cross-dataset experiments while training on UBFC-RPPG

最后,分别以 UBFC-RPPG 和 PURE 作为训练集来测试 BSIPL-RPPG,实验结果如表 4 所示。本文方法 PhysCMAE 的 MeanAE 分别为  $3.96 \text{ min}^{-1}$  和  $6.69 \text{ min}^{-1}$ ,均取得了较优结果。值得注意的是,PhysNet 方法和 PhysFormer 方法虽然分别在跨数据集实验中优于本文方法 PhysCMAE,但它们不能在两个跨数据集实验同时做到优于本文方法。考虑到 BSIPL-RPPG 为自采数据集,其运动模式、受试者分布及采集条件与 UBFC-RPPG、PURE 存在明显差异,更接近现实世界场景。在这种情况下,若模型过度依赖某一训练集中的特定噪声模式或特定心率分布,其性能通常会显著下降。PhysCMAE 使用传统方法构建的 TSTMap 作为正样本,其包含更少的运动噪声。通过对比学习,在特征空间中拉近 MSTMap 与 TSTMap 之间的距离,使模型忽略 MSTMap 中的运动噪声特征,学习与生理信号更加相关的潜在特征,从而表现出更好的泛化性。

图 5 是以 UBFC-RPPG 为训练集在 BSIPL-RPPG 上

进行测试所得到的 Bland-Altman 图,对应表 4 中本文所提出的 PhysCMAE 方法,以及与 MAE (Baseline)、rPPG-MAE 和 PhysFormer 方法的对比结果。在 Bland-Altman 图中,PhysCMAE 在 BSIPL-RPPG 跨数据集测试中具有更集中的误差分布和更小的偏差波动,说明其结果与参考值之间具有较好的一致性。同时结合表 4 可知,PhysCMAE 在面对自采数据这一更真实、更复杂的测试环境时,仍然能够保持较稳定的心率估计性能。这进一步说明,本文方法通过引入生理相关正样本和频率重采样负样本,增强了模型对噪声干扰和场景变化的适应能力。

### 3.4 消融实验

为了探究本文方法 PhysCMAE 中不同模块对最终结果的影响,在 PURE 数据集上进行了心率估计的消融实验研究。

#### 1) 正样本类型消融

为了探究正样本 TSTMap 对 PhysCMAE 结果的影响,除了本文默认使用的 3 种传统方法外,还选择了不同

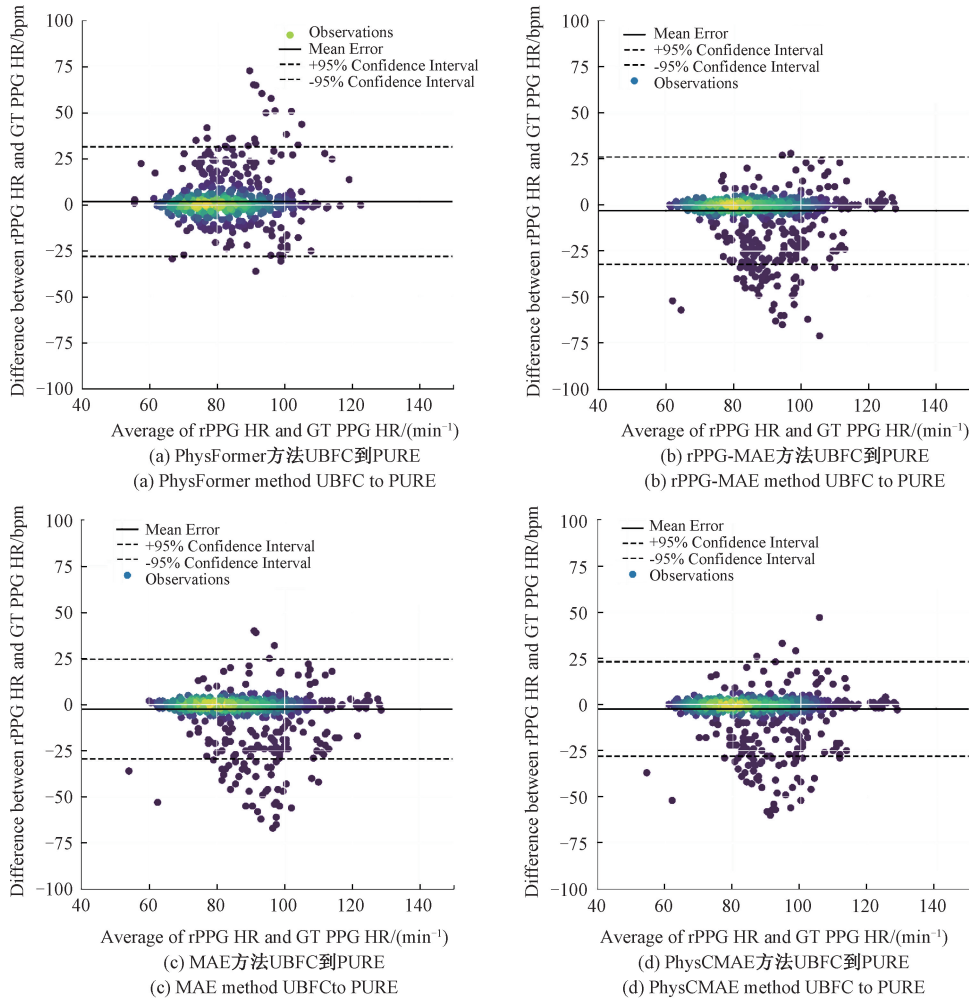


图 5 在 UBFC-RPPG 上训练的 BSIPL-RPPG 跨数据集实验 Bland-Altman 图

Fig. 5 The Bland-Altman figures of BSIPL-RPPG cross-dataset experiments while training on UBFC-RPPG

的传统方法来构建 TMap。其中,本文方法默认使用由 3 种传统方法 CHROM、POS、LGI 构建的 TSTMap 作为正样本,记为 TSTMap-CPL;使用原始 RGB MSTMap 作为正样本,记为 TSTMap-RGB;使用由 CHROM、POS、PBV 3 种传统方法构建的 TMap 作为正样本,记为 TSTMap-CPP;使用由 PCA、ICA、PBV 3 种传统方法构建的 TSTMap 作为正样本,记为 TSTMap-PIP。

实验结果如图 6 所示,可以观察到使用 TSTMap-CPL 作为正样本实现了心率估计的最佳性能。其中,CHROM 利用脉搏信号在不同通道中的相位差和幅度差异,将运动和光照噪声与脉搏信号分离,POS 通过将信号投影到与皮肤色调正交的平面上,消除如光照变化和运动引起的亮度变化,LGI 通过局部群不变性变换将特征映射到对光照变化不敏感的子空间中,并保留对刚性运动不变的脉搏特征。因此由 CHROM、POS 和 LGI 构成的 TSTMap-CPL 能够较全面地保留 rPPG 相关信息。将其作为正样本后,模型在对比学习中被引导去靠近这些更

纯净的生理表示,从而减少对原始 MSTMap 中噪声成分的依赖。

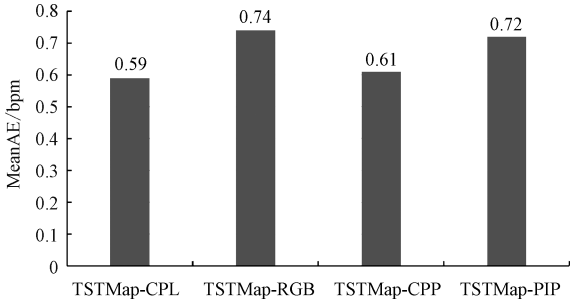


图 6 正样本类型消融实验结果

Fig. 6 Ablation study on positive samples types

2) 分阶段训练策略消融

为了验证引入分阶段训练策略对于协调 CL 与 MAE 二者训练目标冲突的有效性,以及探究引入对比学习的不同轮次对 PhysCMAE 是否有影响,在预训练的不同轮

次  $e_s$  引入对比学习并进行实验,结果如图 7 所示。当在预训练初期就引入对比学习,模型性能明显下降。随着对比学习引入时间逐步后移,性能逐步提升,并在第 80 轮达到最优。而在第 90 轮引入对比学习,性能又有所降低。

该结果表明,MAE 与 CL 虽然具有互补性,但二者并非在任意阶段都能直接协同运作。具体来说,在训练初期,模型尚未充分学习到稳定的底层生理特征,此时过早引入对比约束,将使模型依据更显著但不稳定的噪声差异进行判别,从而干扰重建目标。而当 MAE 已经初步建立起对生理细节的表示能力后引入 CL,就能够在已有生理表征基础上增强实例之间的可区分性,从而实现性能的提升。相反,若对比学习引入过晚,则会因训练时间不足,无法充分发挥判别增强作用。本实验表明,引入分阶段训练策略对于编码器学习具有实例可区分性的潜在生理特征是至关重要的,因此在整个实验中都设定  $e_s$  为第 80 个 epoch。

此外,当不引入 CL 时,性能虽优于基础 MAE 模型,但仍明显弱于完整 PhysCMAE。这表明,数据增强本身可以提供一定帮助,但真正带来性能提升的是基于生理先验的对比约束。

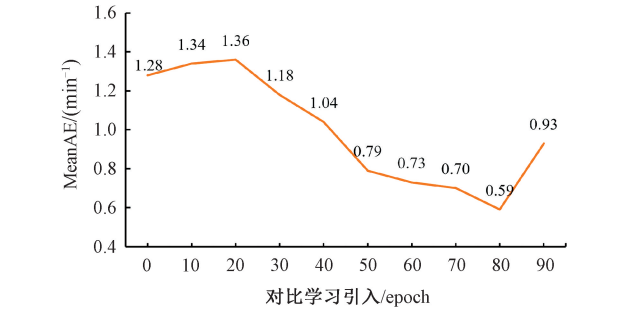


图 7 分阶段训练策略消融实验结果

Fig. 7 Ablation study results of the staged training strategy

3) 少量标签数据微调实验

与 MAE 相同,本文提出的 PhysCMAE 进行无标签自监督预训练,通过标签数据进行微调进行下游心率估计,为了探究在缺少标签数据的情况下本文方法的有效性,进行了少量标签数据微调实验。实验首先在 UBFC 数据集上进行预训练,随后在 PURE 数据集上进行微调。微调数据集和测试集分配与 PURE 数据集内实验相同,对于 PURE 微调数据集的 6 个受试者,每次增加使用一个受试者的数据作为微调数据集来模拟缺少标签数据微调的情况,结果如表 5 所示。

当微调标签极少时,MAE(Baseline)与 PhysCMAE 的性能都较弱,说明仅依赖极少量参考生理信号标签,仍难以建立从预训练生理特征到 BVP 波形的准确映射。表明本文方法虽然降低了标签依赖,但并未完全摆脱下游

心率估计的监督微调需求。然而,随着可用参考生理信号标签的逐步增加,PhysCMAE 始终优于 MAE(Baseline),且在中等标签比例下优势尤为明显。当标签量增至 3/6 时,PhysCMAE 的 MeanAE 从 MAE(Baseline)的 3.13 bpm 下降到 1.11 bpm。这说明,PhysCMAE 预训练阶段所学到的特征具有更高的可迁移性和更好的任务适配性,因此在有限标签监督下能够更快完成对下游心率估计任务的适应。

表 5 少量标签数据微调实验

| Table 5 Fine-tuning experiment with limited labeled data |                                |                                |                              |             |
|----------------------------------------------------------|--------------------------------|--------------------------------|------------------------------|-------------|
| 标签量                                                      | 方法                             | MeanAE<br>(min <sup>-1</sup> ) | RMSE<br>(min <sup>-1</sup> ) | R           |
| 1/6                                                      | MAE (Baseline) <sup>[16]</sup> | 13.14                          | 15.57                        | -0.06       |
|                                                          | PhysCMAE                       | <b>12.35</b>                   | <b>17.48</b>                 | <b>0.22</b> |
| 2/6                                                      | MAE (Baseline) <sup>[16]</sup> | 12.87                          | 16.51                        | 0.26        |
|                                                          | PhysCMAE                       | <b>10.67</b>                   | <b>16.43</b>                 | <b>0.33</b> |
| 3/6                                                      | MAE (Baseline) <sup>[16]</sup> | 3.13                           | 8.42                         | 0.81        |
|                                                          | PhysCMAE                       | <b>1.11</b>                    | <b>3.93</b>                  | <b>0.96</b> |
| 4/6                                                      | MAE (Baseline) <sup>[16]</sup> | 1.93                           | 5.80                         | 0.91        |
|                                                          | PhysCMAE                       | <b>0.94</b>                    | <b>3.57</b>                  | <b>0.97</b> |
| 5/6                                                      | MAE (Baseline) <sup>[16]</sup> | 1.39                           | 5.11                         | 0.93        |
|                                                          | PhysCMAE                       | <b>0.83</b>                    | <b>3.10</b>                  | <b>0.98</b> |
| 6/6                                                      | MAE (Baseline) <sup>[16]</sup> | 1.30                           | 4.72                         | 0.94        |
|                                                          | PhysCMAE                       | <b>0.66</b>                    | <b>1.86</b>                  | <b>0.99</b> |

4 结 论

本文提出一种全新的 rPPG 自监督学习方法 PhysCMAE。该方法将 MAE 与 CL 协同应用于 rPPG 任务,构建了生理信号细粒度特征重建与实例可区分性建模的双重学习机制。同时通过分阶段训练策略确保模型先学习底层生理特征,再增强实例区分能力,有效避免了训练目标冲突。在对比学习模块中,采用传统 rPPG 方法生成正样本以引导模型学习脉搏波共同特征,同时通过频率重采样生成负样本以增强模型对心率频带变化的敏感性,从而提升跨场景泛化能力。后续的实验结果表明 PhysCMAE 在 rPPG 公开数据集和自采数据集上中取得了优越的性能。

然而,PhysCMAE 仍有进一步的改进,未来的研究方向包括以下两点:1) PhysCMAE 仍然使用经过频率增强得到的 FRSTMaps 作为负样本,对于噪声较少数据分布较为均衡的数据集来说,该过程可能引入额外的噪声,造成性能的降低;2) PhysCMAE 目前仍需要一定量的标签数据进行微调才能实现对心率的准确估计,难以适应标签数据极度缺少的情况,因此需进一步探索无标签的 MAE 自监督框架。

参考文献

[1] 张爱华, 常婷婷, 漆宇晟, 等. 基于脉搏信号融合分

- 析的心率监测方法[J]. 电子测量与仪器学报, 2020, 34(10): 163-171.
- Zhang Aihua, Chang Tingting, Qi Yusheng, et al. Heart rate monitoring method based on pulse signal fusion analysis [J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(10): 163-171.
- [2] Chen Xun, Cheng Juan, Song Rencheng, et al. Video-based heart rate measurement: Recent advances and future prospects [J]. IEEE Transactions on Instrumentation and Measurement, 2018, 68(10): 3600-3615.
- [3] 颜文琴, 郑秀娟, 殷中平, 等. 融合面部外观与生理表征的作业人员疲劳判别[J]. 电子测量与仪器学报, 2025, 39(10): 12-21.
- Yan Wenqin, Zheng Xiujian, Yin Zhongping, et al. Fatigue identification of workers by integrating facial appearance and physiological characteristics[J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(10): 12-21.
- [4] Yu Zitong, Li Xiaobai, Zhao Guoying. Facial-video-based physiological signal measurement: Recent advances and affective applications[J]. IEEE Signal Processing Magazine, 2021, 38(6): 50-58.
- [5] Xiang Guoliang, Ou Yuheng, Li Jiaxian, et al. Recovering physiological signals from facial videos: Recent advances and applications in intelligent vehicles [J]. IEEE Transactions on Intelligent Vehicles, 2024, 10(9): 6576-6598.
- [6] Gui Jie, Chen Tuo, Zhang Jing, et al. A survey on self-supervised learning: Algorithms, applications, and future trends[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 9052-9071.
- [7] Chen Ting, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations [C]// International Conference on Machine Learning, 2020: 1597-1607.
- [8] Qian Rui, Meng Tianjian, Gong Boqing, et al. Spatiotemporal contrastive video representation learning [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 6964-6974.
- [9] Gideon J, Stent S. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 3995-4004.
- [10] Sun Zhaodong, Li Xiaobai. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast [C]// European Conference on Computer Vision, 2022: 492-510.
- [11] Sun Zhaodong, Li Xiaobai. Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 5835-5851.
- [12] Yue Zijie, Shi Miaoqing, Ding Shuai. Facial video-based remote physiological measurement via self-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13844-13859.
- [13] Savic M, Zhao Guoying. RS+ rPPG: Robust strongly self-supervised learning for rPPG[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(8): 7911-7924.
- [14] Liu Xin, Zhang Yuting, Yu Zitong, et al. rPPG-MAE: Self-supervised pretraining with masked autoencoders for remote physiological measurements [J]. IEEE Transactions on Multimedia, 2024, 26: 7278-7293.
- [15] Speth J, Van N, Flynn P, et al. Non-contrastive unsupervised learning of physiological signals from video[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 14464-14474.
- [16] He Kaiming, Chen Xinlei, Xie Saining, et al. Masked autoencoders are scalable vision learners [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16000-16009.
- [17] Huang Zhicheng, Jin Xiaojie, Lu Chengze, et al. Contrastive masked autoencoders are stronger vision learners[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 46(4): 2506-2517.
- [18] Shao Hang, Luo Lei, Qian Jianjun, et al. Tranphys: Spatiotemporal masked transformer steered remote photoplethysmography estimation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(4): 3030-3042.
- [19] Niu Xuesong, Yu Zitong, Han Hu, et al. Video-based remote physiological measurement via cross-verified feature disentangling [C]// Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II 16. Springer International Publishing, 2020: 295-310.
- [20] Bulat A, Tzimiropoulos G. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3d facial landmarks) [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 1021-1030.
- [21] De Haan G, Jeanne V. Robust pulse rate from

- chrominance-based rPPG [J]. IEEE Transactions on Biomedical Engineering, 2013, 60(10): 2878-2886.
- [22] Wang Wenjin, Den Brinker A C, Stuijk S, et al. Algorithmic principles of remote PPG [J]. IEEE Transactions on Biomedical Engineering, 2016, 64(7): 1479-1491.
- [23] Pilz C S, Zaunseder S, Krajewski J, et al. Local group invariance for heart rate estimation from face videos in the wild [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018: 1254-1262.
- [24] Bobbia S, Macwan R, Benezeth Y, et al. Unsupervised skin tissue segmentation for remote photoplethysmography [J]. Pattern Recognition Letters, 2019, 124: 82-90.
- [25] 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. IEEE, 2014: 1056-1062.
- [26] Song Rencheng, Chen Huan, Cheng Juan, et al. PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography [J]. IEEE Journal of Biomedical and Health Informatics, 2021, 25(5): 1373-1384.
- [27] Lu Hao, Han Hu, Zhou S K. Dual-gan: Joint bvp and noise modeling for remote physiological measurement [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 12404-12413.
- [28] Špetlík R, Franc V, Matas J. Visual heart rate estimation with convolutional neural network [C]// Proceedings of the British Machine Vision Conference, 2018: 3-6.
- [29] Casado C A, López M B. Face2PPG: An unsupervised pipeline for blood volume pulse extraction from faces[J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(11): 5530-5541.
- [30] Chen Weixuan, McDuff D. Deepphys: Video-based physiological measurement using convolutional attention networks[C]// Proceedings of the European Conference on Computer Vision, 2018: 349-365.
- [31] Yu Zitong, Peng Wei, Li Xiaobai, et al. Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 151-160.
- [32] Yu Zitong, Shen Yuming, Shi Jingang, et al. Physformer: Facial video-based physiological measurement with temporal difference transformer [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 4186-4196.
- [33] Liu Xin, Hill B, Jiang Ziheng, et al. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement [C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023: 5008-5017.
- [34] Yang Jun, Zhu Chen, Wu Renbiao. Robust remote heart rate estimation network based on spatial-temporal-channel learning from facial videos [J]. IEEE Journal of Biomedical and Health Informatics, 2026, 30(2): 1258-1271.
- [35] Zhang Nan, Sun Hongmei, Ma Junrui, et al. A self-supervised learning network for remote heart rate measurement[J]. Measurement, 2024, 228: 114379.

## 作者简介



**王家雄**, 2023 年于合肥工业大学获得学士学位, 现为合肥工业大学硕士研究生, 主要研究方向为智能健康监护。

E-mail: 2023110092@mail.hfut.edu.cn

**Wang Jiaxiong** received his B. Sc. degree from Hefei University of Technology in 2023. Now he is a M. Sc. candidate in Hefei University of Technology. His main research interest includes intelligent health monitoring.



**李佳杰**, 2021 年于扬州大学获得学士学位, 目前正在合肥工业大学攻读仪器科学与技术博士学位。主要研究方向为生物医学信号、图像处理及深度学习。

E-mail: lijiajie@mail.hfut.edu.cn

**Li Jiajie** received his B. Sc. degree from Yangzhou University in 2021. Now he is a Ph. D. candidate in Instrument Science and Technology with the Hefei University of Technology. His main research interests include biomedical signal/image processing and deep learning.



**范伟**, 2018 年于吉林大学获得学士学位, 2023 年于吉林大学获得博士学位, 现为合肥工业大学仪器科学与光电工程学院生物医学工程系副教授, 研究方向为生物医学信号处理。

E-mail: weifan@hfut.edu.cn

**Fan Wei** received his B. Sc. degree from Jilin University in 2018 and Ph. D. degree from Jilin University in 2023. Now he is an associate professor with the Department of Biomedical Engineering, School of Instrument Science and Optoelectronics Engineering, Hefei University of Technology. His main research interest includes biomedical signal processing.



宋仁成, 2005 年于吉林大学获得学士学位, 2010 年于浙江大学获得博士学位, 现为合肥工业大学教授, 主要研究方向为非接触智能感知和情感人机交互。

E-mail: rcsong@hfut.edu.cn

**Song Rencheng** received his B. Sc. degree from Jilin University in 2005, Ph. D. degree from Zhejiang University in 2010. Now he is a professor in Hefei University of Technology. His main research interests include non-contact intelligent perception and human-computer emotional interaction.



刘羽, 2011 年于中国科学技术大学获得学士学位, 2016 年于中国科学技术大学获得博士学位, 现为合肥工业大学教授, 主要研究方向为图像处理、计算机视觉。

E-mail: yuliu@hfut.edu.cn

**Liu Yu** received his B. Sc. degree from

University of Science and Technology of China in 2011, Ph. D. degree from University of Science and Technology of China in 2016. Now he is a professor in Hefei University of Technology. His main research interests include image processing and computer vision.



成娟(通信作者), 2008 年于中国科学技术大学获得学士学位, 2013 年于中国科学技术大学获得博士学位, 现为合肥工业大学教授, 主要研究方向为智能健康监护。

E-mail: chengjuan@hfut.edu.cn

**Cheng Juan** (Corresponding author) received her B. Sc. degree from University of Science and Technology of China in 2008, Ph. D. degree from University of Science and Technology of China in 2013. Now she is a professor in Hefei University of Technology. Her main research interest includes intelligent health monitoring.