基于视频的时空对比无监督的远程生理测量

摘要

基于视频的远程生理测量利用面部视频来测量血容量变化信号，也称为远程光电容积描记术（rPPG）。监督 rPPG 方法需要人脸视频和地面真实生理信号来进行模型训练。本文提出了一种不需要地面真实信号进行训练的无监督 rPPG 测量方法，使用 3DCNN 模型从不同时空位置的每个视频产生多个 rPPG 信号，并用对比损耗训练模型，其中来自同一视频的 rPPG 信号被拉在一起，而来自不同视频的 rPPG 信号被推开。此外，我们引入自相似性 map 改善相位偏差问题，并引用更贴合分布关系的 JS 损失改善在大数据集下不鲁棒的问题。我们在公共数据集上进行测试，结果表明，我们的方法在各数据集上实现了非常接近当前最佳监督 rPPG 方法的精度。

关键词： rPPG；人脸视频；无监督学习；对比学习；自相似性图

1 引言

传统的生理测量需要皮肤接触传感器来测量生理信号，如接触式光电容积描记术（PPG）和心电图（ECG）。然而，皮肤接触生理测量需要特定的生物医学设备，而远程生理测量可以直接使用摄像机记录面部视频，用于测量远程光电容积描记术（rPPG）。基于视频的生理测量只需要现成的摄像机，并且不受物理距离的限制，这对于远程医疗保健 [21,28] 和情感分析应用 [30] 具有巨大的潜力。

在早期的 rPPG 研究 [18,26,27] 中，研究人员提出了手工特征来提取 rPPG 信号。后来，提出了一些基于深度学习（DL）的方法 [4,11,14]，这些方法采用了具有各种网络架构的监督方法来测量 rPPG 信号。基于 DL 的 rPPG 方法需要包括人脸视频和地面真实生理信号的大规模数据集，而获取由接触式传感器测量并与人脸视频同步的地面真实生理信号代价高昂。后来，Gideon 和 Stent [6] 提出了一种自我监督的方法来训练没有标签的 rPPG 测量模型，然而，他们的方法导致了额外的开销，且他们在论文 [6] 中表明，他们的方法容易受到外部周期性噪声的影响。

本文建议使用 3D 卷积神经网络（3DCNN）来处理输入视频，以获得时空 rPPG（ST-rPPG）块。ST-rPPG 块包含沿高度、宽度和时间三个维度的多个 rPPG 信号。根据 rPPG 的时空相似性，我们可以从同一视频的不同时空位置随机采样 rPPG 信号，并将它们拉在一起。根据视频间 rPPG 信号的差异，将不同视频中的 rPPG 信号样本推离。

这项工作的贡献是：1）本文提出了一种 rPPG 表示称为时空 rPPG（ST-rPPG）块，以获得时空维度的 rPPG 信号。2）基于对 rPPG 的观察，包括 rPPG 时空相似性和跨视频 rPPG 相异度，本文提出了一种基于对比学习的无监督方法。3）本文引入了基于自相似性 map 的计

算，改善 rPPG 信号的相位差异。4）本文在三个 rPPG 数据集（PURE [23]、UBFC-rPPG [2]、COHFACE [7]）上进行了实验，本文的方法实现了非常接近监督 rPPG 方法的性能。

2 相关工作

2.1 基于视频的远程生理测量

基于视频的远程生理测量是一种利用摄像头和图像处理技术来监测人体生理参数的方法。Verkruysse 等人 [26] 首次提出 rPPG 可以从绿色通道的人脸视频中测量。早些年提出的大多数 rPPG 方法使用手工程序，不需要数据集进行训练，这被称为传统方法。用于 rPPG 测量的深度学习（DL）方法正在迅速出现。几项研究 [4, 12, 16] 使用具有两个连续视频帧的 2D 卷积神经网络（2DCNN）作为 rPPG 测量的输入。另一种类型的基于 DL 的方法 [14, 15] 使用从不同面部区域提取的时空信号图作为输入，以馈入 2DCNN 模型。最近，提出了基于 3DCNN 的方法 [6, 31] 来实现压缩视频的良好性能。基于 DL 的方法需要人脸视频和地面真实生理信号，因此我们将它们称为监督方法。最近，Gideon 和 Stent [6] 提出了一种无监督的方法来训练没有地面真实生理信号的 DL 模型。然而，他们的方法落后于一些监督方法，并且对外部噪声不鲁棒，且运行速度也不尽如人意。本文的方法有效改善了这些问题。

2.2 对比学习

对比学习是一种广泛用于视频和图像特征嵌入的自我监督学习方法，有助于下游任务训练和小数据集微调 [3, 17]。作为特征提取器的 DL 模型将高维图像/视频映射到低维特征向量。为了训练该 DL 特征提取器，来自同一样本的不同视图（正对）的特征被拉在一起，而来自不同样本的视图（负对）的特征被推开。数据增强（如裁剪、模糊 [3] 和时间采样 [19]）用于获得同一样本的不同视图，以便学习的特征对某些增强是不变的。上面提到的先前工作使用转化学习让 DL 模型为下游任务产生抽象特征，例如图像分类 [3]，视频分类 [19]，人脸识别 [20]。本文的方法有效利用了对比学习的方式，实现了无标签下的对比学习。

2.3 rPPG 发现

本文使用到了 rPPG 的相关性质，包括 rPPG 空间相似性、rPPG 时间相似性、跨视频 rPPG 差异性，这也是设计本文方法和实现无监督学习的前提条件。

rPPG 空间相似性。来自不同面部区域的 rPPG 信号具有相似的波形，它们的 PSD 也相似。一些工作 [9, 10] 也利用 rPPG 空间相似性来设计他们的方法。来自两个不同身体皮肤区域的两个 rPPG 信号之间可能存在小的相位和幅度差异 [8]。然而，当 rPPG 波形被转换成 PSD 时，相位信息被擦除，并且幅度可以被归一化以抵消幅度差异。如图 1所示，来自四个空间区域的 rPPG 波形是相似的，并且它们在 PSD 中具有相同的峰值。

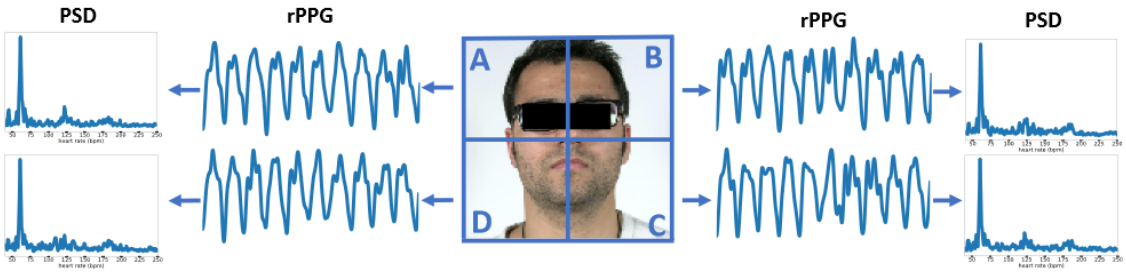


图 1. rPPG 空间相似性图解

rPPG 时间相似性。人力资源不会在短期内快速变化 [6]。斯特里克等人 [23] 还发现，在他们的数据集中，HR 在短时间间隔内略有变化。由于 HR 在 PSD 中有一个主导峰值，PSD 也不会迅速变化。若从一个短的 rPPG 剪辑（例如，10 秒）中随机采样几个小窗口，这些窗口的 PSD 应该是相似的。图 2中显示的是从短的 10s rPPG 信号中采样两个 5s 窗口，并得到这两个窗口的 PSD。可以发现，这两个 PSD 是相似的，并且在相同的频率下具有尖锐的峰值。

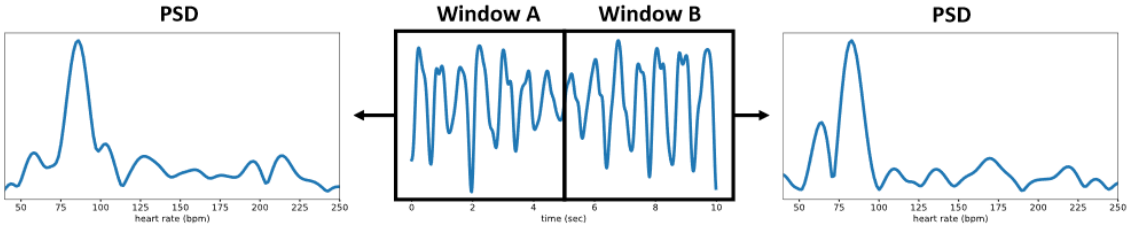


图 2. rPPG 时间相似性图解

跨视频 rPPG 差异。来自不同人脸视频的 rPPG 信号具有不同的 PSD。每个视频都记录了不同的人和不同的生理状态（如运动和情绪状态），因此不同视频的 HR 可能不同 [1]。即使两个视频之间的 HRs 可能相似，PSD 也可能不同，因为 PSD 还包含其他生理因素，如呼吸率和 HRV，这些因素在两个视频之间不太可能完全相同。图 3中显示了 OBFC 数据集中最相似和最不同的跨视频 PSD 对。可以观察到，主要的跨视频 PSD 差异是心率峰值，即由于两个 rPPG 信号表示不同视频下的不同心率，其 PSD 应该不同。

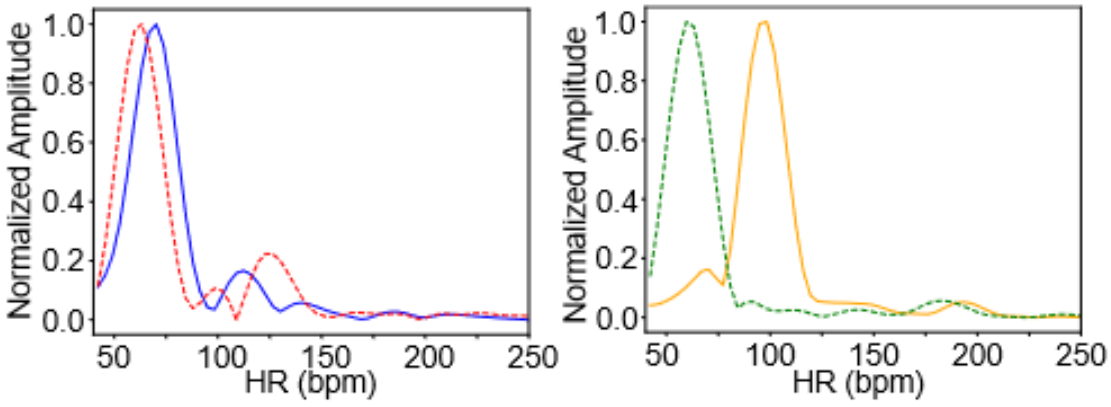


图 3. OBF 数据集中最相似（左）和最不同（右）的跨视频 PSD 对

3 本文方法

3.1 本文方法概述

如图 4所示是本文的方法示意图，本文采用了对比学习的方式，利用 rPPG 信号的三个前提条件（如2.3节中所述）进行对比学习的设计。

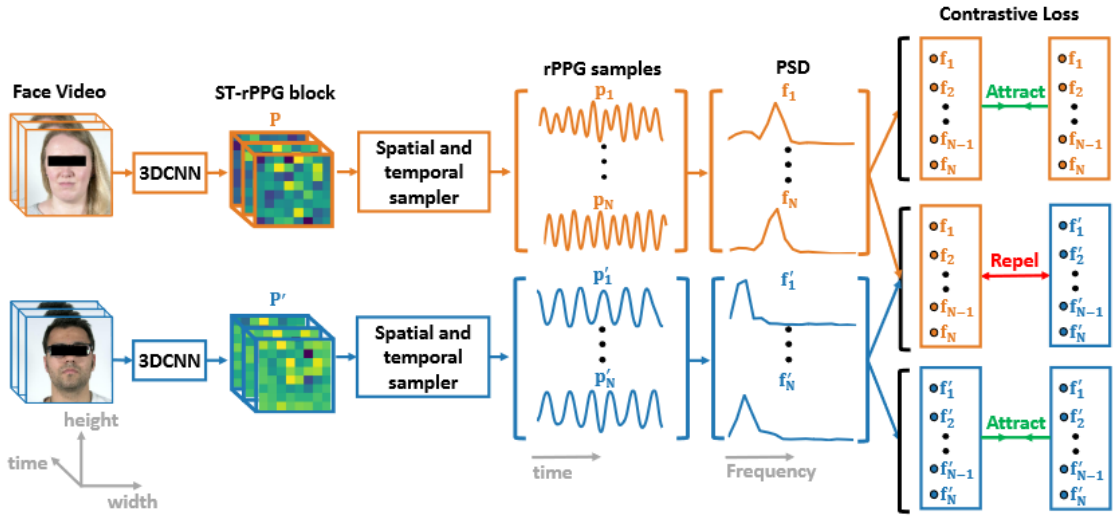


图 4. 方法示意图

首先，本文使用基于 3DCNN 的 PhysNet [29] 修改的模型以获得 ST-rPPG 块表示。修改后的模型有一个形状为 T × 128 ×128 × 3 的 RGB 视频输入，其中 T 是帧数。在模型的最后阶段，本文使用自适应平均池沿空间维度进行下采样，这可以控制输出的空间维度长度。这种修改允许模型输出具有 T × S × S 形状的时空 rPPG 块（具体 ST-rPPG 块描述见3.2节），其中 S 是空间维度长度，如图 4所示。

由于 rPPG 信号具有时间相似性，来自同一视频的 rPPG 信号采样在转换到频率域后得到的 PSD 应该相似，且由于 rPPG 信号具有空间相似性，同一视频下来自不同的面部区域的采样信号的 PSD 也应该相似，因此同一视频下的信号采样可以相互构成正样本对。然而，由于 rPPG 信号在不同视频下的差异性，来自不同视频的采样信号形成的 PSD 相互构成负样本对。基于此，本文通过 ST-rPPG 块采样得到正样本对及负样本对（具体采样器见3.3节），本文最后使用对比学习的方式，将正样本对拉近，负样本对推开，使模型不断学习。

3.2 ST-rPPG 块表示

ST-rPPG 块是时空维度上的 rPPG 信号的集合。本文用 $P \in \mathbb{R}^{T\times S\times S}$ 来表示 ST-rPPG 块。假设在 ST-rPPG 选择一个空间位置 (h, w)。在这种情况下，该位置对应的 rPPG 信号是从原始视频中该空间位置的感受野提取的 $P(\cdot, h, w)$。我们可以推断，当空间维度长度 S 较小时，ST-rPPG 块中的每个空间位置都具有较大的感受野。ST-rPPG 块中每个空间位置的感受野可以覆盖面部区域的一部分，这意味着 ST-rPPG 块中的所有空间位置都可以包含 rPPG 信息。

3.3 时空 rPPG（ST-rPPG）块采样器

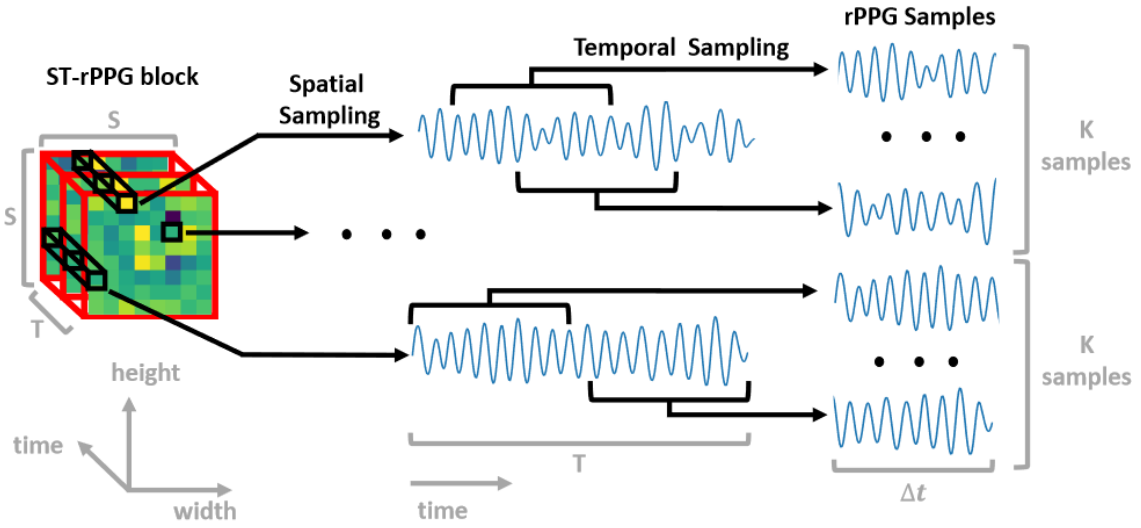


图 5. 时空采样器

如图 5所示，从 ST-rPPG 块采样几个 rPPG 信号。对于空间采样，我们可以在一个空间位置获得 rPPG 信号 $P(\cdot, h, w)$。对于时间采样，我们可以从 $P(\cdot, h, w)$ 中采样一个短的时间间隔，最终的时空样本是 $P(t \to t + \Delta t, h, w)$ 其中 h 和 w 是空间位置，t 是开始时间，Δt 是时间间隔长度。对于一个 ST-rPPG 块，我们将循环所有空间位置，并对 K 个 rPPG 剪辑进行采样，每个空间位置随机选择开始时间 t。因此，我们可以从 ST-rPPG 块中得到 S×S×K 个 rPPG 片段。在我们的模型被训练并用于测试后，我们可以直接在空间维度上平均 ST-rPPG，以获得最终的 rPPG 信号。

3.4 损失函数定义

如图 4 所示，本文方法将从数据集中随机选择两个不同的视频作为输入。对于一个视频，可以得到一个 ST-rPPG 块 P，一组 rPPG 样本 $[p_1, \ldots, p_N]$ 和相应的 PSD$[f_1, \ldots, f_N]$。对于另一个视频，可以得到一个 ST-rPPG 块 P'，一组 rPPG 样本 $[p'_1, \ldots, p'_N]$ 和相应的 PSD$[f'_1, \ldots, f'_N]$。如图 4 右部分所示，对比损耗的原理是将来自同一视频的 PSD 拉在一起，并将来自不同视频的 PSD 推开。注意，根据以秒为单位的 HR 范围限制，仅使用 0.66 Hz 和 4.16 Hz 之间的 PSD（这是由 Hr 波动范围决定的）。

3.4.1 正损失项

正损失项。根据 rPPG 时空相似性，我们可以得出结论，来自同一 ST-rPPG 块的时空采样的 PSDs 应该是相似的。我们可以使用均方误差作为损失函数，从同一视频中提取 PSD（正对）。正损耗项 L_p 如公式 1所示，它用正对的总数进行归一化。

$$L_p = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} (\|f_i - f_j\|^2 + \|f'_i - f'_j\|^2) / (2N(N-1)) \tag{1}$$

3.4.2 负损失项

据跨视频 rPPG 的不同，我们可以得出结论，来自两个不同的 ST-rPPG 块的时空采样的 PSD 应该是不同的。因此，本文方法使用负均方误差作为损失函数，从两个不同的视频中推开 PSD（负对）。负损耗项 L_n 如公式 2所示，它用负对的总数进行归一化。

$$L_n = -\sum_{i=1}^{N}\sum_{j=1}^{N}(\|f_i - f_j'\|^2)/N^2 \tag{2}$$

总损失函数为 $L = L_p + L_n$，它是正负损失项之和。

4 复现细节

4.1 与已有开源代码对比

此部分为必填内容。如果没有参考任何相关源代码，请在此明确申明。如果复现过程中引用参考了任何其他人发布的代码，请列出所有引用代码并详细描述使用情况。同时应在此部分突出你自己的工作，包括创新增量、显著改进或者新功能等，应该有足够差异和优势来证明你的工作量与技术贡献。

复现论文作者提供了复现代码学习，因此直接从作者的 github 仓库中下载方法进行运行。代码中包括实现的整体框架及训练、测试代码，但不包括预处理整体方法。数据预处理主要包括视频数据转帧数据并使用对应的人脸识别模型获得人脸框并截取保存，并对 ground truth 按频率采集并插值，其应与视频帧率对应。其中，COHFACE 数据集视频需要调整帧率以适应 ground truth。

复现结果与论文中相近，但从实验中发现方法中存在的不足之初，并加以改进，本节将详细描述其不足之处并提出改进的方法。

4.1.1 改进动机

在复现中发现原文方法仍然存在一定的不足，分别是模型在非理想场景中的数据集下的不适用性及由相位差导致的结果预测不准确性。

首先是非理想场景中的数据集的不适用性，在实验中测试结果如表 1 所示，可以发现原方案在理想场景下的数据集（UBFC、PURE）中可以实现较好的效果，但却在非理想条件数据集（COHFACE）下产生较差的效果。因为虽然均方差 loss 能产生一个不错的效果，但其更多是数值上的拉近，而不是一个分布的对齐，所以在复杂的数据集上效果不佳。由此，我们可以使用分布对齐的 JS 损失来改善。

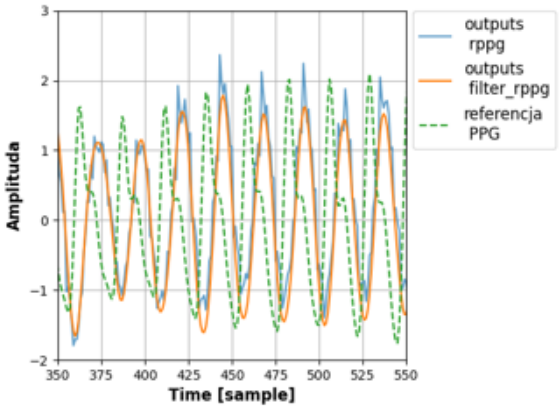


图 6. UBFC 测试结果与 ground truth 的相位偏差图

同时，如图 6所示，在测试时会发现数据波形存在误差，例如我们测试的值和 ground truth 会存在一定的偏差，因为指尖传递的信息和面部反应的结果会有时间差，这将导致测试结果的不准确性。因此，在测试的时候应该使用相似性图 SSM 来优化，以确保测试的准确性。该方法虽然不是正向改善模型，却在准确指标测量上有重要意义。

4.1.2 改进方法

首先，改进后的方式采用了 JS 损失来弥补原方案下 MSE 损失的不足，优化拉近分布间的差距，JS 损失由 JSD(Jensen–Shannon divergence) [5] 得到，如公式 3所示，

$$JS(p,q) = (KL(p\|q) + KL(q\|p))/2 \tag{3}$$

其中散度测度 $KL(\cdot)$ 表示 Kullback-Leibler（KL）散度 [25]。

由此，可以得到 JS 损失的正损失和负损失，分别如公式 4和公式 5所示。

$$L_{p_{js}} = \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}(JS(f_i,f_j) + JS(f_i',f_j'))/(2N(N-1)) \tag{4}$$

$$L_{n_{js}} = -\sum_{i=1}^{N}\sum_{j=1}^{N}(JS(f_i,f_j'))/N^2 \tag{5}$$

则最终的总损失改变为 $L = L_p + L_n + \alpha(L_{p_{js}} + L_{n_{js}})$，其中 α 是超参数，实验中设置为 0.1。

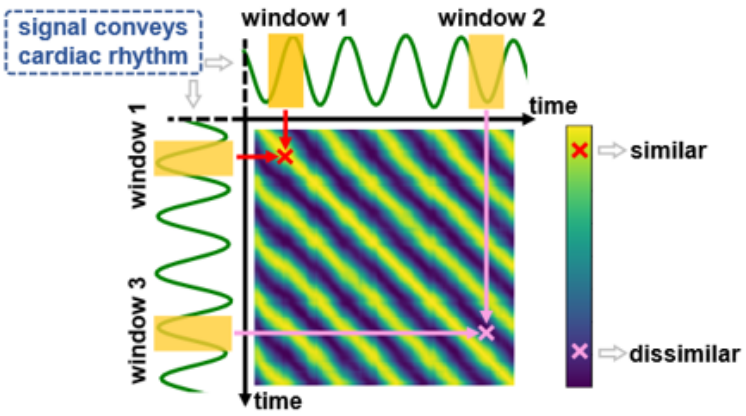


图 7. 自相似生理图

同时，本文引入自相似性生理图，通过窗口与窗口之间的相似性形成自相似性图，根据自相似性图的值求得新的测试波形，从而减少相位带来的影响，使得预测结果更加准确。如图 7所示是典型的自相似生理图，其中较高的值 (亮) 表示两个窗口之间的时间消息相似，反之亦然。设定固定窗口后，将窗口与窗口之间的值绘制为自相似性图。对应的每条斜对角线的内容表示的是相同时间间隔下的窗口之间的相似性值。

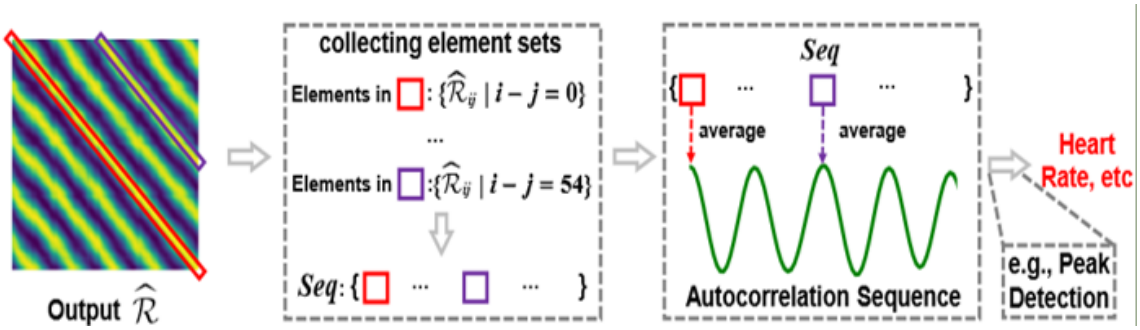


图 8. 由自相似生理图计算心率流程图

如图 8所示是由自相似生理图计算心率的过程，将各斜对角线的值求均值可以得到对应的波形，由波形进行后处理可以获得对应的心率值。这种方式得到结果的好处是，消除了相位偏差的影响，引入该模块，在准确指标测量及提升鲁棒性等方面具有重要意义。

4.2 实验数据及参数环境设置

4.2.1 数据集

本文测试了三个常用的 rPPG 数据集，涵盖在不同场景下录制的 RGB 视频。数据集内部测试使用 PURE [23]、UBFC-rPPG [2]、COHFACE [7] 用于跨数据集测试。PURE 有 10 个受试者在 6 种不同设置下录制的面部视频，包括稳定任务和运动任务，运动任务中受试者会左右频繁转头，拍摄环境下无自然光照影响。UBFC-rPPG 包含 42 名受试者的面部视频，拍摄环境为较理性环境，无自然灯光影响，无偏头等情况。COHFACE 包含 40 位受试者录制的 160 个 videos，拍摄条件存在自然光照影响，包含有噪声的数据集及无噪声的数据集。

4.2.2 实验设置

我们仿照的方式对各数据集进行了相同的处理，其中，我们使用 MTCNN 方式对视频帧逐帧获取人脸框，将人脸按 1.2:1 的扩大边界框裁剪出，并调整为 128×128 分辨率。对于时空采样器，我们在灵敏度分析部分评估了 ST-rPPG 区块的不同空间分辨率和时间长度。根据结果，我们在其他实验中固定了如下参数。我们设置 K = 4，即在 ST-rPPG 块的每个空间位置随机选择四个 rPPG 样本。我们设置 ST-rPPG 块的空间分辨率为 2 × 2，ST-rPPG 块的时间长度为 10 秒。每个 rPPG 样本的时间间隔 t 是 ST-rPPG 块时间长度的一半。我们使用 AdamW 优化器训练模型，学习率为 10-5，训练迭代次数为 30 轮。在每次训练迭代中，输入分别是来自两个不同视频的两个 10 秒钟片段。测试时，我们将每个测试视频分成不重叠的 30 秒片段，并计算每个片段的 rPPG。我们找到 rPPG 信号 PSD 的最高峰来计算 HR。

4.2.3 评价指标

根据之前的研究 [15,31]，我们使用平均绝对误差（MAE）、均方根误差（RMSE）和皮尔逊相关系数（R）来评估心率测量的准确性。对于 MAE、RMSE，数值越小误差越小，而对于 R，数值越大误差越小。

4.3 创新点

基于时空对比无监督的远程生理测量使用对比学习的方式进行无监督学习，该方式与传统方式不同，无需标签数据，且结果优于大部分其他方法，甚至比有监督方法的结果更优。新的方案引用了合理的方式改善了原文结果存在的问题。第一，引用了 JS 损失的方式，结合 MSE 创建新的损失函数，根据 PSD 分布之间更准确的关系拉近或推开正、负样本对，产生了优于原文结果中更好的效果。第二，使用了自相似性图 SSM 进行结果改进，使得实验结果更准确且使人信服。第三，使用了动态可视化的方式展示了 rPPG 信号的作用及其心率测量的有效性。

5 实验结果分析

5.1 数据集内及跨数据集测试

实验中，我们对三个不同的数据集采用多种其他的经典的监督学习和无监督学习方法进行对比。监督学习方式包括 Physnet [29] 及 Dual-GAN [13]，无监督学习方式包括 Gideon2021 [6] 及 Contrast-phys [24]，我们的方法改进了 Contrast-phys 方式的不足，提高了其在非理性环境下的鲁棒性。具体数据集内测试实验结果如表 1所示，各数据集下最好的结果加粗显示，第二的结果用下划线显示。

如表 1 所示，在数据集的对比下，无监督的 Contrast-phys 可以实现较好的效果，甚至在部分数据集如 PURE 上优于监督学习方法 Dual-GAN [13]。同时，对比无监督学习，该方法优于无监督学习 Gideon2021 [6]，甚至结果在 MAE 指标上提高一半以上。

但从实验结果中可以发现，Contrast-phys [24] 在非理想条件下表现很差，如在存在头部运动的 PURE 数据集中，以及在存在光照不均匀的 CPHFACE 数据集中，都表现较差，其中 COHFACE 尤为明显。因为传统的方式在数值上做拉近已经不再适合，我们需要考虑更优

表 1. **数据集内测试对比**

Method types	Method	UBFC			PURE			COHFACE			Average
		MAE	RMSE	R	MAE	RMSE	R	MAE	RMSE	R	
Supervised	Physnet [29]	0.60	1.93	0.83	1.67	5.42	0.44	17.62	21.67	0.16	1.80
	Dual-Gan [13]	**0.44**	**0.67**	**0.99**	0.82	1.31	0.99	-	-	-	-
Unsupervised	Gideon 2021 [6]	1.85	1.22	0.99	2.3	2.9	0.99	-	-	-	-
	Contrast-phys [24]	0.64	1.00	0.99	1.00	1.40	0.99	13.63	17.16	0.03	1.00
	Updated Contrast-phys	0.58	1.30	0.99	**0.34**	**0.84**	**0.99**	**11.88**	**12.18**	**0.31**	**0.94**

化的分布对齐的方式，并改善其相位差带来的影响。而通过本文方法改进后的 Contrast-phys 模型，不论是在 PURE 还是 COHFACE 上都表现不错，在 COHFACE 上相比较原来的方法提高了 2，在 PURE 上也从原来的基础上提升了 50%。由于其方式主要提升在非理性条件下的性能，因此在较理想环境下的 UBFC 的实验结果提升较小，但仍有所提升。

本文还测试了跨数据集性能，本文采用在 UBFC 数据集上训练的模型对 PURE 数据集进行测试，如表 2所示是测试的结果。

表 2. **跨数据集测试**

Method	MAE	RMSE	R
Physnet [29]	1.5	3.87	0.96
SiNC-rPPG [22]	3.81	-	0.87
Contrast-phys [24]	1.17	2.73	0.97
Updated Contrast-phys	**1.05**	**2.54**	**0.97**

从表 2 数据中可以看出，我们的方式在跨数据集测试中也可以实现最好的效果，且相较于监督学习的方式，本文效果明显更优，在 MAE 及 RMSE 指标上均表现提升近 30%。由此可以看出，无论是在数据集内测试还是跨数据集测试，本文的方法都可以达到最好的性能。

5.2 鲁棒性测试

本文仿照 Gideon2021 [6] 中的方式对 PURE 及 UBFC 做鲁棒性显著图测试，具体来说，我们使用最后一层的每个通道的结果，并使用其置信度进行加权，其原理和热力图类似，因为 rPPG 信号主要表现在 RGB 的绿色通道中，因此显示绿色通道的结果。图 9 显示了两种

情况下的显著性图：1）手动注入周期性噪声，即在视频的左上角注入类似 rPPG 信号的周期性干扰噪声；2）涉及视频中头部运动帧，本文选取了带有头部运动伪影的视频片段进行测试本文方法的鲁棒性。

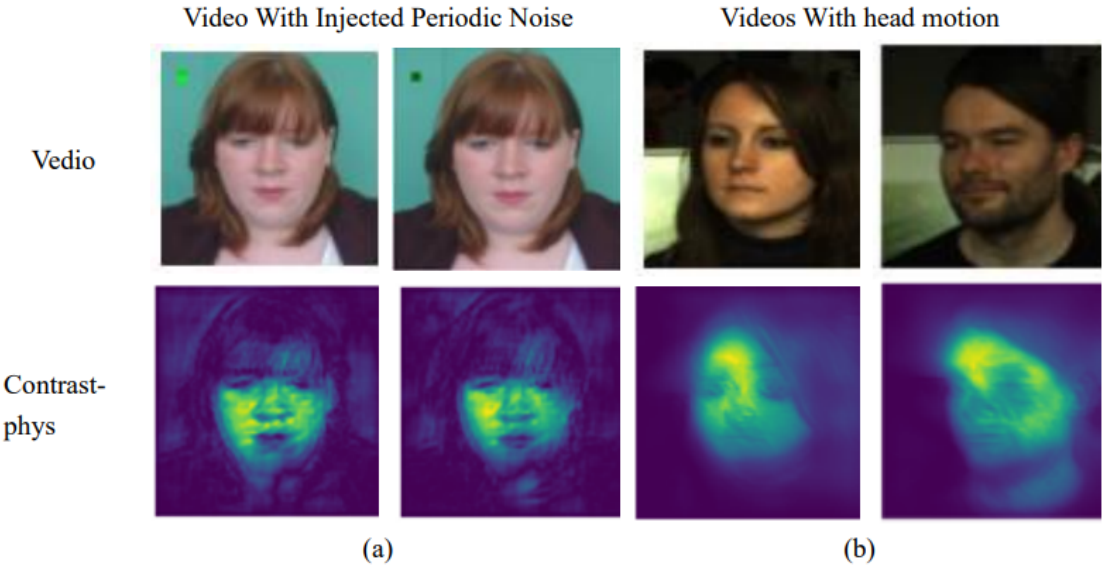


图 9. 显著性图鲁棒性测试

图 9(a) 显示在视频的左上角注入了周期性噪声时，我们的方法不会受到噪声的干扰，仍能聚焦于皮肤区域。我们的方法对噪声具有鲁棒性，因为噪声只存在于一个区域，而这违反了 rPPG 的空间相似性，我们的方法关注到了空间相似性，故可以改变局部噪声的影响。图 9(b) 显示了涉及头部运动时的显著性图。我们的方法的显著性图聚焦并激活了大部分皮肤区域，即使在头部运动出现伪影等情况下，仍然覆盖了大部分的面部区域。体现本文方法在视频数据人脸晃动下仍然具有不错的性能。

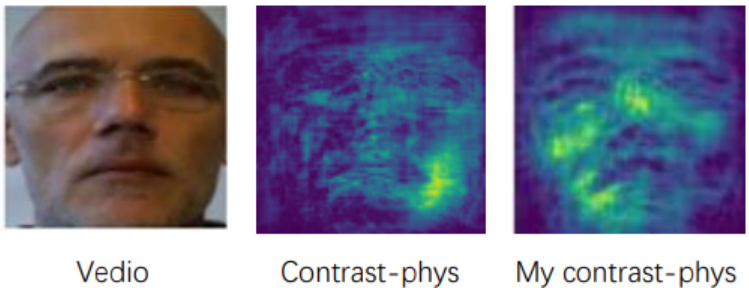


图 10. COHFACE 数据集非理想环境下测试

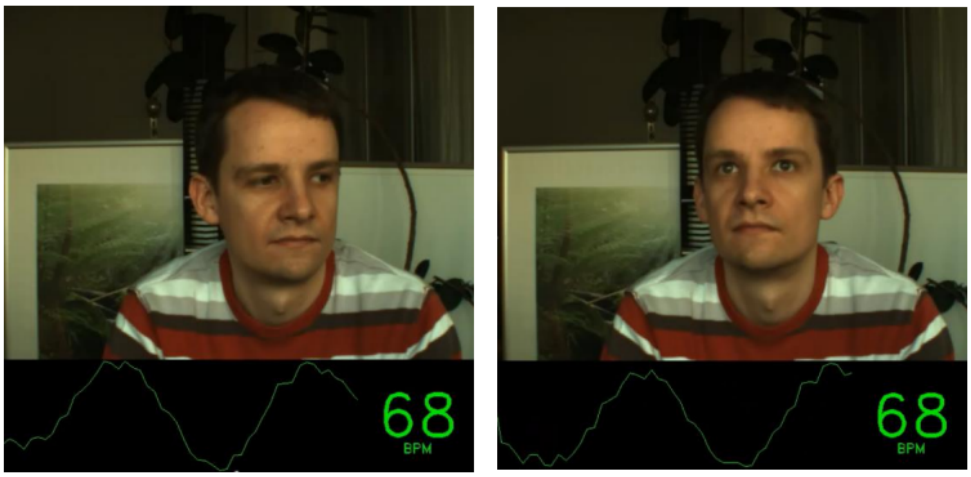
同时，本文对于非理想条件下的 COHFACE 数据集进行测试，该数据集的拍摄视频处于自然非均匀光照影响下，且数据集中包含现实世界中存在的各种噪声，如光照噪声，摄像噪声等。如图 10所示，可以发现原 Contrast-phys 在该数据集下测试结果非常差，甚至出现无法识别面部区域的情况，而本文的改进方法使其无论从轮廓还是从显著性表示上，都有一个更优的性能。因为本文的方法是基于分布对齐进行的，比简单的数值损失约束更有效，更加鲁棒。

由鲁棒性测试可以看出，本文方法在视频场景下具有较好的鲁棒性，无论是出现环境噪

声还是信号噪声的情况下，都能较好的应对。同时，当视频出现比较常见的面部运动、人物运动情况下，本文方法仍然能表现优异。

5.3 HR 指标稳定性测试

rPPG 任务主要对标生理指标检测，本文采用心率 HR 指标测试来检测模型是否具有一定的稳定性。具体来说，本文使用 PURE 数据集中具有头部运动的数据进行测试，按照正常的思想下，在较短帧数内的波形不会出现较大突变，即 HR 在波形不突变的情况下不应该出现跳变。实验中随机选择视频帧序列进行测试，如图 11 所示是其中某两帧的结果，波形展示的是该帧及其前后 60 帧的信号波形。



(a) turn head left (b) head up

图 11. 心率测试

由图 11可以看出，我们的方法在 HR 上表现较稳定，不会出现明显跳变及波动，其稳定性也可以通过 4.2 中实验结果的 RSME 指标体现，其值越小，即表示我们的结果表现越稳定。由此可总结，不管从什么角度看，我们的结果都是既准确又稳定的。

5.4 消融实验

表 3. **消融实验结果**

JSP	SSM	MAE	RMSE	R
w/o	w/o	13.63	17.16	0.03
w/o	w.	13.2	15.59	0.15
w.	w/o	**11.66**	14.1	0.28
w.	**w.**	11.88	**12.18**	**0.31**

本文对优化的方法进行了消融实验，具体包括对 JS 损失及自相似性图模块的实验，如表 3所示是在非理性环境数据集 COHFACE 上进行的消融实验的结果。我们对是否使用 JS

损失及是否使用自相似性模块对结果的影响做了比较，其中“w/o”表示不适用该方法，“w.”表示加入该方法后的结果。从实验结果可以看出，使用 JSP 损失可以有效提升整体新能，而使用自相似性图模块虽然提升较小，但却在稳定性指标 RMSE 及相关性 R 上表现不错，因此，本文方法的结合使得提升效果更为明显，缺一不可。

6 总结与展望

rPPG（remote photoplethysmography）是一种通过无需接触皮肤的方式来检测心率和其他生理信号的技术。它基于对人的脸部进行视频或图像处理，利用视频图像中微小的皮肤颜色变化来提取心率相关的生理信号。这种技术可以应用于健康监测、情绪识别、压力检测等领域，并且在个人健康管理和医疗诊断中具有潜在的应用前景。

本文利用无监督的对比学习方式，替代了传统的需要标签监督的方式，该方法可以在没有标签监督的情况下进行训练，有效解决了目前缺少有标签数据训练的问题，并实现了精确的 rPPG 测量。本文的方法基于对 rPPG 的三个观察结果，并利用时空对比度实现无监督学习，除此之外，本文就模型存在的问题进行有效改进，引入分布对齐 JSP 损失以拉近信号分布距离，引入自相似性图模块以解决相位偏差问题。本文进行了丰富的实验分析，从数据集内测试到跨数据集测试，还加入了鲁棒性测试等，从实验结果可以看出，本文方法在各个方面均明显优于之前的无监督基线 [6]，且对比先前的无监督方式有显著的提升。

在此，我将本文贡献总结为以下几点：（1）引用无监督对比学习方式，在没有标签监督下进行训练；（2）引入分布对齐损失以拉近信号分布距离；（3）引入自相似性模块以解决相位偏差问题；（4）本文方法在目前大多数无监督方法乃至监督学习方法中都表现更优异的性能，其中比监督学习 [2] 提升 30%，比无监督学习 [6] 提升 70%。

需要注意的是，虽然 rPPG 技术具有许多潜在的优势，但在实际应用中仍然需要考虑到光照条件、运动伪影、噪声干扰等因素对信号提取的影响，以及隐私和数据安全等问题。在未来的工作中，我们希望将多考虑其他因素的干扰情况，尤其是人为视频运动带来的伪影问题，除此之外，目前我们正在研究 rPPG 的跨域训测以及持续学习方式，并引入了更深入的分布距离度量，敬请期待。

参考文献

[1] All about heart rate (pulse). https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure/all-about-heart-rate-pulse.

[2] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365, 2018.

[5] Bent Fuglede and Flemming Topsoe. Jensen-shannon divergence and hilbert space embedding. In *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE, 2004.

[6] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3995–4004, 2021.

[7] Guillaume Heusch, André Anjos, and Sébastien Marcel. A reproducible study on remote heart rate measurement. *arXiv preprint arXiv:1709.00962*, 2017.

[8] Alexei A Kamshilin, Victor Teplov, Ervin Nippolainen, Serguei Miridonov, and Rashid Giniatullin. Variability of microcirculation detected by blood pulsation imaging. *PloS one*, 8(2):e57117, 2013.

[9] Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal. Distanceppg: Robust non-contact vital signs monitoring using a camera. *Biomedical optics express*, 6(5):1565–1588, 2015.

[10] Antony Lam and Yoshinori Kuno. Robust heart rate measurement from video using select random patches. In *Proceedings of the IEEE international conference on computer vision*, pages 3640–3648, 2015.

[11] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 392–409. Springer, 2020.

[12] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020.

[13] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12404–12413, 2021.

[14] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019.

[15] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 295–310. Springer, 2020.

[16] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. The benefit of distraction: Denoising camera-based physiological measurements using inverse attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4955–4964, 2021.

[17] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[18] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010.

[19] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[21] Jingang Shi, Iman Alikhani, Xiaobai Li, Zitong Yu, Tapio Seppänen, and Guoying Zhao. Atrial fibrillation detection from face videos by fusing subtle variations. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2781–2795, 2019.

[22] Jeremy Speth, Nathan Vance, Patrick Flynn, and Adam Czajka. Non-contrastive unsupervised learning of physiological signals from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14464–14474, 2023.

[23] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014.

[24] Zhaodong Sun and Xiaobai Li. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022.

[25] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[26] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

[27] Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016.

[28] Bryan P Yan, William HS Lai, Christy KY Chan, Stephen Chun-Hin Chan, Lok-Hei Chan, Ka-Ming Lam, Ho-Wang Lau, Chak-Ming Ng, Lok-Yin Tai, Kin-Wai Yip, et al. Contact-free screening of atrial fibrillation by a smartphone using facial pulsatile photoplethysmographic signals. *Journal of the American Heart Association*, 7(8):e008585, 2018.

[29] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019.

[30] Zitong Yu, Xiaobai Li, and Guoying Zhao. Facial-video-based physiological signal measurement: Recent advances and affective applications. *IEEE Signal Processing Magazine*, 38(6):50–58, 2021.

[31] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 151–160, 2019.