Dual-GAN：用于远程生理测量的联合BVP和噪声建模

**摘要：**

基于生理测量的远程光电容积脉搏波描记法（remote photoplethysmography, rPPG）在健康监测、情感分析等领域具有很大的应用价值。现有的方法大都关注于如何从人脸视频中加强或提取非常微弱的血容量脉冲（blood volume pulse, BVP）信号，但很少明确地对主导面部视频内容的噪声进行建模。所以这些方法的在一些前所未见的场景中，可能会表现出很差的泛化性能。本文提出了一种新颖的对抗学习方法，用于基于rPPG的生理测量，该方法使用对偶学习生成对抗网络（Dual Generative Adversarial Networks, Dual-GAN）对BVP预测因子和噪声分布进行联合建模。BVP-GAN旨在学习一种从输入到真实BVP信号的抗噪声的映射，Noise-GAN旨在学习噪声分布。这两个GAN可以增强彼此能力，从而促进BVP和噪声之间的特征解缠。此外，我们提出了一种名为ROI 对齐融合（ROI alignment and fusion, ROI-AF）块的即插即用模块，以抑制不同ROI之间的非一致性，同时，在ROI方面，挖掘来自于更广阔的感受野的信息特征。与最先进的方法相比，我们的方法在基于面部视频的心率、心率变异性和呼吸频率估计方面取得了更好的性能。

1. 介绍

如心率（heart rate, HR）、呼吸频率（respiration frequency, RF）和心率变异性（heart rate variability, HRV）的生理信号时人类健康状态的重要指标。传统上，这些生理信号是使用心电图 (ECG) 和光电容积脉搏波描记法 (PPG) 测量的；两者都是基于皮肤接触的方法，但上述侵入性方法可能会导致人类不适。最近，基于远程光电容积脉搏波描记法（rPPG）的非接触式生理测量方法[18, 19, 25]受到了越来越多关注，并且大多数方法都可以使用商品相机录制的面部视频来进行基于rPPG的生理测量。基于rPPG的生理测量背后的原理是皮肤的光吸收随着心跳引起的周期性血容量变化而进行周期性地变化。因此，如果我们能够捕捉到周期性的肤色变化，我们就可以获得心率。然而，这种肤色变化非常微弱，易受光照、头部运动等各种噪声的影响。

基于rPPG的生理测量的早期方法通常使用 PCA 或 ICA 来分解原始时间信号 [1、20、28] 或执行颜色空间转换来提取 BVP 信号，如 CHROM [8] 和 POS [37] 方法。这些手工制作的算法通常是对背景噪声做出一定的假设，例如，运动对不同颜色通道的强度变化具有相同的影响，在此基础上，CHROM 方法可以通过计算三个通道之间的比例来消除运动影响[8]. 然而，由于人脸不是一个理想的朗伯体，所以这种假设并不总是适用于每个面部区域。 因此，虽然手工制作的方法可能不需要训练并且具有相对较好的泛化能力，但生理测量精度仍有很大的提升空间。此外，即便有大型数据集，手工制作的方法也可能无法有效地利用大型训练数据集来学习信息特征。

随着深度学习 (deep learning, DL) 在各种计算机视觉任务[14、23、24]中的巨大成功，DL 方法也被研究用于基于 rPPG 的生理测量 [25、27]。 考虑到面部视频中 BVP 信号的 PSNR 较低，DL 方法通常在使用卷积神经网络 (convolutional neural networks, CNN) 学习信息特征之前首先计算时空图 ( 时空图) [25] 或帧差 (DOF) [27] 用于生理信号。然而，大多数 DL 方法只关注于如何从视频中提取 BVP 信号，而忽略了主导视频内容的背景噪声的建模。

一些研究表明，使用带有人工噪声的合成生理信号可以有益于基于 DL 的生理测量方法的训练 [21、27]。 然而，这些合成噪声通常是通过将高斯噪声与三角函数（如正弦或余弦）或阶跃脉冲混合来编译的。显然，这种合成生理信号生成方法无法在实际场景中复制真实的噪声分布。 因此，它们提高模型鲁棒性的有效性也受到来限制。

为了解决这些问题，我们提出了一种新颖的对抗学习方法，通过使用对偶学习生成对抗网络（Dual Generative Adversarial Networks, Dual-GAN）同时对 BVP 预测因子和噪声分布进行建模，以实现基于鲁棒 rPPG 的远程生理测量。如图 1 所示，我们首先从输入视频计算时空图 (spatial-temporal map, 时空图)，以获得 BVP 信号的初步表示。然后，我们使用一个 GAN 模型来学习从时空图到 BVP 的映射，其中生成器（也称为 BVP 估计器）旨在生成与真实 BVP 相似的 BVP 信号，而鉴别器旨在区分生成的 BVP 和真实的 BVP 信号。 另一个 GAN 用于对关于BVP的噪声分布进行建模。其中生成器（包括一个无噪声时空图和一个纯噪声时空图生成器）旨在生成与从视频计算的时空图相似的合成时空图，而鉴别器（共享）旨在区分合成 时空图和真实 ST地图。 Dual-GAN 可以增强彼此的能力。 前者提高了后者的噪声分布学习能力，而后者则像在线数据增强一样，可以反过来提高前者 BVP 预测器的抗不可见噪声的鲁棒性。

此外，现有的方法在卷积过程中不加区别地处理不同 ROI（时空图的各行）的时间信号；然而人脸并不是理想的朗伯体，不同 ROI 的时间信号应该有不同的 BVP 和噪声分布。因此，我们还提出了一种名为ROI 对齐融合（ROI alignment and fusion, ROI-AF）块的即插即用模块，它可以按ROI进行卷积来处理这种各ROI之间的噪声分布不一致问题，并从更广的感受野融合 BVP 特征。

这项工作的贡献如下：

1）我们提出了一种基于 rPPG 的生理测量的新型对偶GAN 架构，它不仅可以对BVP 预测器进行建模，还可以通过对抗学习显式建模噪声分布，从而可以获得更稳健的，抗不可见噪声的 BVP 表示。

2) 我们提出了一个即插即用的 ROI-AF 块，它可以在传统的卷积层之后使用，以解决不同 ROI 之间的噪声和 BVP 分布不一致的问题。

3）所提出的方法在数据集内和跨数据集测试下均优于最新方法的 HR、HRV 和 RF 估计，再次展示其再复杂场景下的鲁棒性。

1. 相关工作
   1. 远程生理测量

远程生理测量旨在从商品摄像头录制的视频中实现HR、HRV、RF估计。传统方法主要基于某一特定的皮肤反射模型，通过进行信号分解来获得BVP信号。De Haan 等人提出了一种基于颜色空间投影（color space projection, CHROM)的色度，来消除头部运动对HR估计带来的影响。后来，提出了一种逐像素的CHROM方法来进一步改进HR估计[38]。Wang等人提出了基于三个颜色通道的相关性的空域子空间旋转（spatial subspace rotation, 2SR），以提高HR估计的鲁棒性。De Haan等人研究了不同波长下的rPPG信号的特征，提出了一种BVP矢量方法来提取脉冲信号[9]。从盲信号分离的角度出发，采用独立分量分析（Independent component analysis, ICA）将BVP信号从三个颜色通道中分离出来，用于HR估计[28]。这些方法是在特定假设下手动设计的，当图像采集条件改变时，可能不再有效。

深度学习（Deep learning, DL）模型具有强大的非线性拟合能力，已成功应用于各种计算机视觉任务。 也有人尝试研究基于DL的远程HR估计[6,15,21,25,32,39,42]。DeepPhys计算帧间差异并使用深度卷积网络提取生理信号。还提出了一种注意机制来减少运动的影响 [6]。 Wang 等人提出了一种新颖的两步 CNN，采用低秩约束损失来获得可靠的特征 [39]。 一个 3D 卷积网络（命名为 Spatial-Temporal Net）被用来通过融合空间和时间信息直接从视频估计 HR [32]。除了 BVP 信号之外，人脸视频还包含许多不相关的信息，因此一些研究试图为生理信号设计有效的手工表示 [15, 25]。 有研究将时频表示作为 CNN 的输入，它直接累加从视频中提取的生理信号的频率分量 [15, 29]。 有研究设计时空表示作为CNN的输入，它是从面部不同 ROI 中提取的时间生理信号的组合 [25]。 还有一些研究关注如何抑制与生理信号混合在一起的噪声[18, 26]。 Niu 等人试图通过交叉验证进行特征解缠来去除 MSTMap 中的噪声 [26]。 Lee 等人尝试使用元学习来应对模型部署过程中的噪声分布变化[18]。Song 等提出通过 GAN 学习 BVP 信号分布以提高 BVP 波形质量。大多数现有方法都关注如何从视频中增强和提取 BVP 信号，但忽略了对背景噪声的显式建模，这对于改进 BVP 信号提取也很重要。 与现有方法不同，我们建议通过对抗性学习同时对 BVP 预测器和噪声分布进行建模。因此，我们可以更好地分离 BVP 特征和噪声特征，以实现更鲁棒的生理测量。

2.2. 深度生成式噪声建模

生成模型旨在从有限数量的数据中学习真实的数据分布。使用深度神经网络对这些模型进行参数化的最新进展已成功应用于各种计算机任务，如图像转换 [44]、面部编辑 [7]、图像超分辨率 [10] 和图像去噪 [4]。最常用和最有效的深度生成模型之一是生成对抗网络（GAN）[12]。 GAN 可以有效地使用无监督学习来学习未知数据分布，并在许多应用中取得了巨大的成功。

例如，GAN 已被广泛用于生成无噪声图像 [5、16、43]。 对于远程生理测量任务，一些方法尝试使用带有噪声的合成生理信号进行数据增强并提高模型的鲁棒性 [21, 27]。 这些合成生理信号是通过将三角函数（如正弦或余弦）和步进脉冲等干净信号与高斯噪声混合而生成的，这些信号可能无法很好地复制真实世界的噪声。 考虑到 GAN 在噪声建模方面的有效性，一些研究调查了基于 GAN 的生理信号生成，并报告说比以前的生理测量方法更好的性能 [11、30]。

我们提出了一种用于联合建模噪声分布和生理估计器的Dual-GAN，以进行远程生理测量。我们的方法虽基于 GAN，但与现有方法相比有几个显著差异：

(i) 与 [21、27] 中的噪声合成相比，我们的模型可以通过 GAN 学习真实生理信号的分布，因此可以生成更多逼真的，可以在实际应用中复制的生理信号

(ii) 与图像去噪方法 [5、16、43] 不同，我们的方法不需要配对的数据（如噪声和无噪声）来学习噪声分布。

(iii) 不同于 [11, 30] 中的合成方法，我们同时学习 BVP 估计器和生成噪声模型，这可以增强彼此的能力，从而更好地实现BVP 和噪声信号之间的特征解缠。

1. 提出的方法

我们将第i个输入视频表示为vi，将对应的真实BVP信号表示为s i gt。远程生理信号测量的目的是学习一个映射：F : vi→sigt。因为BVP信号与背景人脸内容（表征身份和属性）和视频中存在的噪声相比十分微弱，现有的方法通常是通过借助手工制作的变换[8，28，37]或深度学习[6，18，42]为F构建复合函数来提取BVP信号。然而，这些方法没有明确地对噪声分布进行建模。因此，BVP信号和背景噪声之间的解缠可能是次优的，导致在新场景中的泛化能力较差。我们提出了一种对抗性学习框架，使用Dual-GAN网络对BVP预测器和噪声分布进行联合建模。如图1所示，除了通过由生成器（名为BVP估计器Fb）和鉴别器D组成的GAN学习从STMap mi到BVP si gt的映射之外，我们还通过由两个子生成器Gphy和Gnoise组成的同级GAN以及鉴别器D学习噪声分布，其中D由两个GAN共享。我们方法的细节描述如下。

3.1 时空图

正如在[6.25]中所讨论的，直接将CNN应用在人脸视频上可能无法有效地挖掘生理信号信息。因此，我们像[22,25]一样，选择使用STMap作为CNN的输入，这通过丢弃大部分不相关的背景内容来建立初步的生理信号表示。用mi表示从vi计算得来的STMap，mi的维度为n×l×c，其中n表示ROI的数量，l表示一个切片的帧数，c表示RGB三个通道。然后，我们的目标是建立一个映射Fb 从mi→sigt

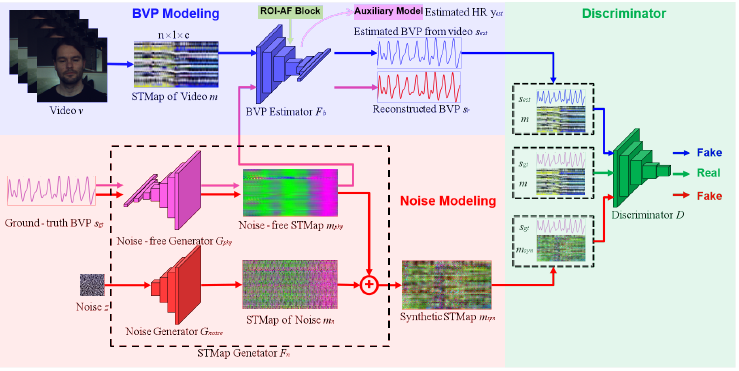
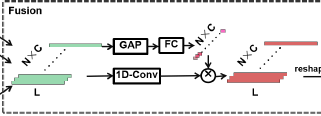


图1用于对BVP预测器和噪声分布联合建模的Dual-GAN的架构。对BVP建模的GAN包含一个从STMap映射至BVP信号的生成器Fb，一个用来区分预测BVP信号s est和真实的BVP信号sgt的判别器D。用于噪声建模的GAN包含两路生成器，由真实BVP和随机噪声变量生成STMap，并且使用一个公用的判别器D来区分合成的STMapms和真实的STMap m。Dual-GAN架构允许我们对噪声进行间接监督，因此我们可以更好地对噪声分布进行建模，并且对BVP信号实现更好地特征解缠。

3.2 ROI对齐融合模块。

STMap的每行mi表示人脸上一个ROI的原始时域信号。因为人脸不是一个完美的朗伯体，所以不同ROI的BVP和噪声分布是不同的。然而，现有的方法通常是对不同ROI的时域信号进行相同的卷积。当滤除不同ROI中的噪声以获得BVP信号时，这种方式可能不是最佳的。更合理的方法是对不同的ROI进行不同的卷积。从另一方面来说，如[17,26]中显示的那样，从面部区域不同的ROI中提取的BVP信号应是几乎同步的，并且，一个相对于ROI更大的感受野对提取高PSNR的周期性的BVP信号很有帮助。然而，由于卷积核感受野的限制，目前的CNN在每次卷积操作中，只能利用到一点相邻的ROI。

为了解决这些问题，我们提出了一种简单但有效的模块，名为ROI对齐融合模块，来进行跨ROI的特征对齐和融合。如图2所示，一个CNN的特征图或输入时空图是按行划分的，每行进行一维卷积，以此来减轻BVP和噪声分布（在不同ROI之间的）差异，并达到跨ROI特征对齐的目的。然后，对齐的特征会按通道堆叠，通过带通道注意力机制的一维卷积，即全局池化加两层线性层来融合来自于各ROI的对齐的特征。最后，融合的特征图再重塑成原来的形状。提出的ROI-AF模块实际是一个即插即用，并能插入在CNN网络的不同卷积快中。



（针对时空图用常规的对图像的卷积似乎不太合适，理解：时空图相较于一般的图片具有特殊性。本来的局部卷积是将相邻的几个ROI融合提取特征，近几帧(l）哪个通道(c)更重要？，不太能考虑全局， 这样fusion之后，能从全局看哪个通道更重要，所以是对全局的融合）

3.3. DualGAN

Dual-GAN 通过对抗学习对 BVP 预测器和噪声分布进行联合建模。下面，我们详尽阐述BVP-GAN和Noise-GAN。

BVP-GAN：

BVP-GAN由一个生成器Fb和一个判别器D组成。Fb学习从时空图m到真实BVP信号sgt的分布，即sest=Fb(m)。 D旨在区分预估的BVP信号sest和真实的BVP信号sgt。如图3(a)所示，生成器Fb由四个ConvBlock和四个UpBlock组成，它将时空图作为输入，输出一个预估的BVP信号。四个ROI-AF模块被插入在四个卷积模块前，来减少BVP和噪声在不同ROI之间的差异。除了BVP信号的预测，我们还引入了一个副主任吴，即从时空图预测HR的回归任务，旨在通过多任务学习提升生成器 Fb 特征学习鲁棒性。

**我们仅关注其周期性，而非幅度变化。**因此，我们选用预估BVP sest和真实BVP sgt之间的负皮尔逊相关损失来衡量预测BVP信号sest的质量：



其中，PCor(·)表示皮尔逊相关，我们还用一个频域损失，定义为预估BVP的频谱分布和真实BVP的独热码ogt的交叉熵损失：



其中PSD(·)表示sest的频谱，CE(·)表示交叉熵损失。真实HRygt也可以用一个独热向量表示ogt=[0,…,0,1,0,…]，且’1’表示ygt对应的索引。PSD(sest)可以被看做是HR的独热吗形式，如p = [0:1,…, 0:1, 0:5,0:1; …]。对于辅助HR回归任务，我们有监督地计算预估HRyest和真实HRygt之间的L1距离。总之，生成器Fb的总体损失函数可以写作：



（这个损失很巧妙我感觉，因为是一个向量所以能用相关损失衡量，除了用对抗还有这一优化的途径。）

其中，λ1λ2和λ3是平衡参数，在我们的实验中，我们按各损失大小根据经验经验设定λ1=0.2，λ2=1，λ3=0.1。

BVP-GAN的判别器D包含5各卷积核，一个GAP和一个FC，这将STMap和BVP信号作为输入，输出真或假，如图3(d)所示。D的损失函数可以定义为：

手机截图图人的照片上写着字

中度可信度描述已自动生成

其中，Fn(·)表示Noise-Gan的生成器，旨在从真实BVP信号和一个随机的噪声变量生成一个时空图，将在下一小节中讨论。通过生成器Fb和判别器D的生成学习，我们期望Fb能够预估尽可能逼真的BVP信号。

**Noise-GAN**：在BVP预估中，使得现有方法不能对噪声分布进行显式建模的显著问题是没有配对数据（有噪声数据和无噪声数据）以供监督学习。我们通过使用一个非监督的方法来解决这个问题，如监督已知BVP分布和未知噪声分布之和。

具体地，Noise-GAN 抱哈按一个时空图生成器Fn用来从BVPsigna和随机噪声变量到STMap的生成，和一个于BVP-GAN共享的判别器D。生成器Fn使用一个两路结构，包括两个子生成器（Gphy和Gnoise）和一个总的操作。Gphy旨在从真实BVP信号生成一个无噪声的时空图，即mphy=Gphy(sgt)，Gnoise旨在从一个服从高斯分布的随机噪声变量z生成一个噪声时空图，即mn=Gnoise(z) 我们假定一个将噪声和BVp信号加在一起的模型，将mphy和mn相加获得了一个合成时空图msyn。所以，Fn旨在生成一个合成的时空图msyn，使其尽可能地逼近BVP-GAN使用的真实时空图m。

如图片3(b)(c)所示，Gphy包含4各ConvBlock和四个UpBlock，Gnoise包含四个UpBlock。Fn的对抗损失也是方程4.

我们的Dual-GAN结构也能允许我们引入额外的监督来确保各部分能够按我们的期望奏效。具体的，我们期望生成器Fb在BVP-GAN中保留有用的BVP信息时，只要求其滤除噪声。然后，基于Fb和Gphy，期望重构的BVPsignal sr 和sgt是相同的，，这样的限制可以通过最小化夫皮尔逊相关系数Lr

来实现：



训练策略。我们使用一个可选的训练策略来训练Dual-GAN。在每个batch中，我们对真是寻来你数据（m，sgt和ygt）和由我们的模型生成的合成数据（msyn sest）进行采样。这些数据使用如下：

（i）训练Fb使用训练数据（m，sgt和ygt）和合成数据msyn，和对应的标签sgt和ygt，使用损失函数Eq3.

（ii）训练Gphy，使用损失函数Eq5的Lr。

（iii）训练Fb和Gnoise，通过最小化Eq4的L joint

（iv）训练D，将所有生成器参数固定，通过最大化Eq4中的Ljoint

鉴于提出的Dual-Gan架构和可选的训练策略，期望 和 都能获得真实的BVP信号sgt。然后，Gnoise改进的噪声建模能力将会反过来提升Fb的抵抗未在训练集出现的噪声鲁棒性。

这可以提升未知场景下所提出方法的生理测量的泛化能力。

一些方法也基于Dual-GAN，例如[13,40,41]，但是我们的工作与他们的不同：

（i）用在[13]的Dual-GAN用于新的视网膜和分割图像合成，这需要成对的数据（掩模和视网膜图像）进行监督；然而我们方法中的噪声建模不需要成对的无噪声数据以供监督。

（ii） [41]使用Dual-GAN执行风格级图像转换，但不保留像素级对应关系。我们的BVP特征与带有噪声的STMap分离需要精确的时间对齐。

（iii）在[40]中使用了Dual-GAN，通过知识融合重用多个预先训练的网络进行多标签分类，这与我们的特征解缠是完全不同的任务。

1. 实验结果