Continual Learning for Remote Physiological Measurement: Minimize Forgetting and Simplify Inference

Qian Liang, Yan Chen, and Yang Hu*

University of Science and Technology of China
qianliang@mail.ustc.edu.cn, {eecyan,eeyhu}@ustc.edu.cn

Abstract. Remote photoplethysmography (rPPG) has gained significant attention in recent years for its ability to extract physiological signals from facial videos. While existing rPPG measurement methods have shown satisfactory performance in intra-dataset and cross-dataset scenarios, they often overlook the incremental learning scenario, where training data is presented sequentially, resulting in the issue of catastrophic forgetting. Meanwhile, most existing class incremental learning approaches are unsuitable for rPPG measurement. In this paper, we present a novel method named ADDP to tackle continual learning for rPPG measurement. We first employ adapter to efficiently finetune the model on new tasks. Then we design domain prototypes that are more applicable to rPPG signal regression than commonly used class prototypes. Based on these prototypes, we propose a feature augmentation strategy to consolidate the past knowledge and an inference simplification strategy to convert potentially forgotten tasks into familiar ones for the model. To evaluate ADDP and enable fair comparisons, we create the first continual learning protocol for rPPG measurement. Comprehensive experiments demonstrate the effectiveness of our method for rPPG continual learning. Source code is available at https://github.com/MayYoY/rPPGDIL.

Keywords: Remote physiological measurement · Continual learning · Domain prototype

1 Introduction

The measurement of physiological signals, heart rate (HR) and heart rate variability (HRV) for instance, holds significant importance across various fields, such as medical diagnosis and emotion recognition [5,35,58]. Traditional physiological measurement relies on specialized contact sensors, which are not only inconvenient to use but also potentially uncomfortable for subjects. In contrast, remote photoplethysmography (rPPG) can extract physiological signals associated with heartbeats by detecting periodic color changes in facial skin. Due to its capability of facilitating contactless physiological measurements using a common camera, rPPG has gained growing attention in recent years [10, 13, 30, 32, 33, 38, 41].

* Corresponding author

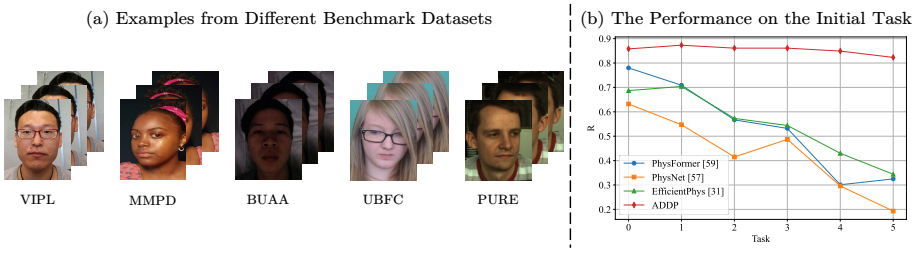

Fig. 1: (a) Example sequences from popular benchmark rPPG datasets. Samples from different datasets have different backgrounds, skin colors, head motions, lighting conditions, *etc*. (b) The performance of mainstream rPPG methods on the initial task during the incremental learning process. Previous methods exhibit obvious catastrophic forgetting while our method (ADDP) effectively alleviates this phenomenon.

Although existing deep learning-based rPPG approaches [9,30,38,57,59] have achieved impressive performance, most of them only focus on the setting where training data is collected in a single session and the model is trained only once. However, developing a robust model for real-world applications necessitates a large and diverse dataset that encompasses various scenarios. Collecting such a dataset in a single session can be extremely challenging and nearly impossible. In practice, data from diverse scenarios is often collected gradually and the model is updated accordingly each time a new set of training data becomes available. This setting of training a model sequentially on a series of datasets (referred to as "tasks" interchangeably in this paper) is known as continual learning. Due to its great practical value, continual learning has attracted significant attention recently. However, this setting has not been explored by the rPPG researches.

In the context of rPPG measurement, factors such as lighting condition, skin color and motion can lead to distribution shifts across different datasets (See Fig. 1 (a)). Consequently, the sequential training on a series of rPPG datasets can be framed as domain incremental learning, which is the primary focus of this work. The key challenge of incremental learning is catastrophic forgetting [15,44], where the model tends to forget the past knowledge after finetuning on new tasks, resulting in a significant decrease in performance on previous data. To investigate whether this phenomenon occurs in rPPG measurement, we conduct a continual learning evaluation and Fig. 1 (b) shows the performance of three state-of-the-art rPPG methods on the initial task after learning new tasks sequentially. It can be observed that these methods experience significant performance degradation during the continual learning process, indicating the models' struggle to retain past knowledge and the occurrence of catastrophic forgetting.

To alleviate catastrophic forgetting, many methods preserve samples from old tasks and replay them during the training on new tasks. Although these methods demonstrate impressive performance, applying them to rPPG measurement is not practical. RPPG methods typically require long videos as input, with clip lengths exceeding 100 frames. In such cases, replaying previous samples could re-

sult in an extremely heavy storage burden. Furthermore, the stored facial videos involve subjects' privacy information, which is undesirable in real-world applications. By contrast, rehearsal-free algorithms retain no samples from previous tasks and are more practical. These methods commonly relies on class prototypes to preserve decision boundary [34, 61, 62], which is infeasible for rPPG measurement as both of the ground truth heart rate and rPPG signal are continuous. Another promising approach involves leveraging the generalizability of pre-trained models and training task-specific prompts [47, 52, 53]. However, our empirical results demonstrate that prompt-based methods are not directly applicable for rPPG measurement (refer to Sec. 4.1 and Sec. 5.3 for details). To circumvent these challenges, we propose a novel rehearsal-free method based on **AD**apter and **D**omain **P**rototypes (**ADDP**) as follows:

- Adapter-based Finetuning: In order to improve the model's stability and adapt it to new tasks more efficiently, we utilize adapter [20] to finetune the frozen backbone. Additionally, we design a Difference Normalization (DiffNorm) module for the backbone, which can effectively integrate dynamic and appearance features of the input videos.
- Prototype-based Augmentation: Although adapter finetuning minimizes the changes of parameters, continually finetuning a single group of adapters inevitably leads to forgetting. For consolidating the past knowledge, we propose to extract domain prototypes, including style prototypes and noise prototypes, of previous tasks and employ them to augment new samples during the training on new tasks.
- Prototype-based Inference Simplification: Drawing inspiration from the fact that people tend to simplify unfamiliar tasks by relating them with familiar ones, we propose to utilize the style prototype most familiar to the model to transfer the style of test samples, which enables the model to solve the inference problem in its most proficient manner.

The contributions of this work are 1) To the best of our knowledge, our approach is the first to explore domain incremental learning for rPPG measurement. 2) To tackle this problem, we first employ adapter to finetune the backbone. Additionally, we design domain prototypes to replace class prototypes for such regression problem and introduce prototype-based augmentation to alleviate catastrophic forgetting. Furthermore, we leverage the style prototypes to simplify the inference. 3) We establish a practical domain incremental learning benchmark for rPPG measurement. Extensive experiments show the superiority of the proposed method.

2 Related work

2.1 Remote Physiological Measurement

Remote physiological measurement relies on detecting color changes in facial skin caused by heartbeats to estimate vital signals like heart rate. Traditional methods mainly leverage blind source separation or color space transformations to extract rPPG signals that possess a high signal-to-noise ratio (SNR) [10,11,41,51].

However, these methods heavily rely on prior assumptions and struggle to perform well in complex environments. Recently, deep learning approaches have been successfully used in rPPG measurement [9, 30, 38, 57, 59]. They utilize 3D-CNN or modified vision transformer (ViT) [12] to extract physiological features from raw videos [57, 59] or carefully designed STMaps [37, 38]. Additionally, considering that the distribution shifts between the testing and training data will limit the performance of models, some researchers have explored the problems of domain generalization and domain adaptation for rPPG measurement [13, 33]. Unfortunately, these models still fail to generalize or adapt well to challenging domains, further prompting us to address the domain shifts in rPPG measurement through domain incremental learning.

2.2 Continual Learing

The primary challenge of continual learning is catastrophic forgetting. Existing methods can be broadly categorized into three classes. Architecture-based methods assign specific parameters for each task through learnable masks or expandable modules to mitigate inter-task interference [26, 45, 46, 56]. Regularization-based methods penalize changes to important parameters or model predictions to balance the old and new tasks [1, 2, 8, 18, 25]. Rehearsal-based methods store a subset of samples from old tasks in a buffer and replay them during the training on new tasks to consolidate past knowledge [3, 4, 19, 43].

Since rehearsal with stored data of old tasks will incur a large storage cost and violate data privacy, more and more researchers have started focusing on rehearsal-free continual learning. Apart from regularization and architecture-based methods, some recent approaches opt to extract feature centroids as class prototypes and utilize augmented prototypes to maintain decision boundary of previous tasks [34, 62]. Additionally, the emerging parameter efficient finetuning-based (PEFT) methods [14, 52–54] have demonstrated the effectiveness of frozen backbones with PEFT modules for forgetting minimization. S-Prompts [52] trains task-specific prompts for each task to assign distinct subspace for different domains to tackle domain incremental learning. It further employs feature centroids to select proper prompts for inference. LAE [14] designs a unified class incremental learning framework with PEFT modules. It calibrates the adaptation speed of PEFT modules relative to the classifiers and leverages predicted logits to ensemble an online model and an offline model. In this work, we specifically focus on domain incremental learning for rPPG measurement. Drawing inspiration from PEFT and prototype-based methods, we propose a novel and rehearsal-free method based on adapter finetuning and label-irrelevant domain prototypes that are more suitable for rPPG measurement.

3 Preliminaries

3.1 Problem Formulation

In this work, we focus on domain incremental learning for rPPG measurement (rPPG DIL). Different from existing rPPG works, the model sequentially learns

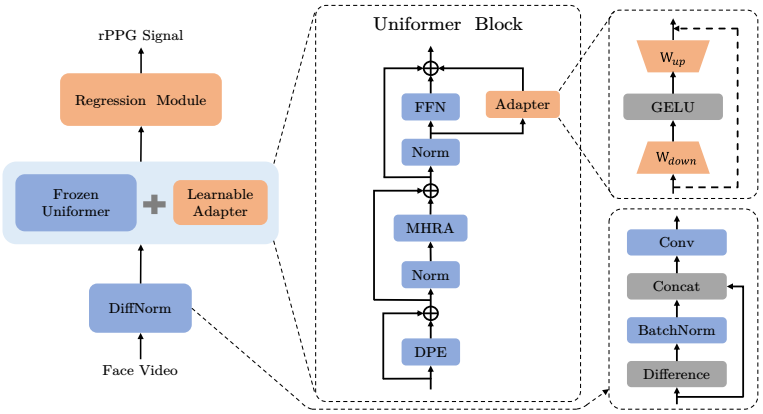

Fig. 2: The architecture of our model. The base model is Uniformer which can extract the crucial local features for rPPG measurement. Our DiffNorm module can effectively fuse the appearance and dynamic features. Only the adapter and regression modules (the orange blocks) are learnable after the initial task.

knowledge on different tasks with distribution shifts and is expected to perform well on all the tasks. Formally, the model $f(\cdot;\boldsymbol{\theta},\boldsymbol{\phi})$ is trained on tasks $\mathcal{T} := \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n\}$ with multiple domains in sequence, where $\boldsymbol{\theta}, \boldsymbol{\phi}$ are parameters of the feature extractor and the regression module, respectively. During the inference stage, given an input facial video $\boldsymbol{x} \in \mathbb{R}^{3\times T\times H\times W}$, the model is required to accurately predict the ground truth rPPG signal $\boldsymbol{y} \in \mathbb{R}^T$ corresponding to $\boldsymbol{x}$ without knowing which task the test sample belongs to. Note that variations in illumination, subject's skin color, head motion, noises introduced by video compression and many other factors can lead to distribution shifts across rPPG datasets, and the primary challenge is to mitigate the catastrophic forgetting caused by domain shifts.

3.2 Uniformer and Adapter

In this part, we introduce two major parts of our feature extractor.

Uniformer [28] is a strong spatial-temporal model for video classification. It consists of 4 stages and each stage contains a stack of Uniformer blocks. As shown in Fig. 2, a Uniformer block mainly comprises three modules: Dynamic Position Embedding (DPE), Multi-Head Relation Aggregrator (MHRA) and Feed Forward Network (FFN). DPE leverages 3D depthwise convolution to encode position information and is friendly to arbitrary input lengths. FFN is a Multi-Layer Perceptron (MLP) same as the one used in vanilla ViT. The key module, MHRA, unifies 3D convolution and self-attention in a concise transformer format. It utilizes convolution in the first two stages to reduce spatial-temporal redundancy and capture local features while employing self-attention to model global dependency in the subsequent stages.

Adapter [20], one of the most commonly used PEFT modules, can efficiently finetune a large model by inserting a small module to any layer of it. As illustrated in Fig. 2, an adapter consists of a pair of projection matrices and a non-linear activation function (GELU in this work). The forward procedure can be formulated as follows:

$$\boldsymbol{x}' = \boldsymbol{x} + \mathbf{W}_{up}(\mathrm{GELU}(\mathbf{W}_{down}(\boldsymbol{x}))) \tag{1}$$

where $\boldsymbol{x}$ is the input of the adapter. The skip connection is optional and we do not employ it in this work.

4 Methodology

4.1 Overall Framework

The architecture of our model is shown in Fig. 2. Following S-Prompts [52], we adopt the learning paradigm of freezing a pre-trained backbone and tuning a small number of learnable parameters with PEFT. In Sec. 5.3, we empirically validate that vanilla ViT is not suitable for our problem and that local features are crucial for rPPG measurement. Therefore, we turn to the aforementioned Uniformer [28]. However, the prompt finetuning [27] used by S-Prompts is not directly applicable due to the convolution layers in Uniformer.[1] To this end, we employ the adapter finetuning and attach an adapter in parallel with the FFN for each Uniformer block. To better leverage the expanded training data and facilitate knowledge sharing among tasks, we continually finetune a single group of adapters rather than following S-Prompts to separate different PEFT modules for different tasks. Moreover, inspired by EfficientPhys [31], we design a module called DiffNorm at the input layer, as illustrated in Fig. 2. Building upon the normalization module of EfficientPhys, this module further utilizes convolution and concatenation to effectively integrate the appearance and dynamic features of the input videos:

$$\boldsymbol{x}' = \mathrm{Conv}([\boldsymbol{x}; \mathrm{BN}(\mathrm{Diff}(\boldsymbol{x}))]) \tag{2}$$

where x is the input video, $[\cdot;\cdot], \mathrm{BN}(\cdot), \mathrm{Diff}(\cdot)$ are the concatenate, batch normalization and difference operations, respectively and $\boldsymbol{x}'$ will be the input of the Uniformer. In summary, the feature extractor can be divided into two parts: the backbone that consists of the Diffnorm Module and the Uniformer, and the adapter for finetuning. The backbone is only trained on the initial task. Then we freeze it and only finetune the adapter and the regression module for the following tasks.

To address catastrophic forgetting, we propose two strategies as illustrated in Fig. 3. Firstly, during model finetuning, we utilize prototype-based augmentation to consolidate the past knowledge. Secondly, during testing, the prototype-based inference simplification is used to convert tasks that may have been forgotten or are challenging into a form that the model is most familiar with.

[1] It is feasible but less effective for rPPG measurement to forcefully utilize prefix to finetune the last two stages of Uniformer. See the supplementary material for details.

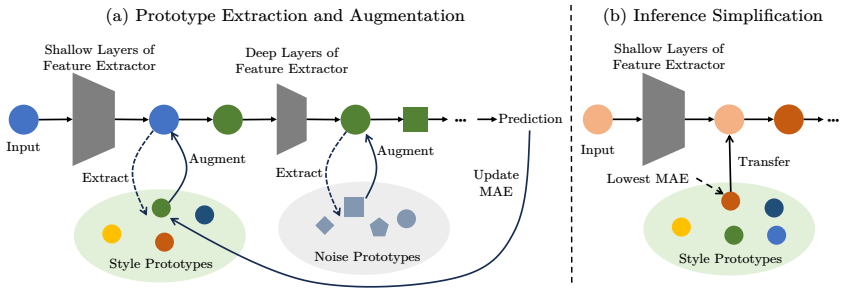

Fig. 3: Overview of our domain prototype-based strategies. (a) During training stage, two types of domain prototypes, *i.e.* style prototypes and noise prototypes, are extracted for each task. Meanwhile, we randomly select domain prototypes of previous tasks to augment the training samples and record the training MAE of the selected style prototypes. (b) In the inference stage, the style prototype with lowest MAE is selected to transfer test samples into a style that can be easily processed.

4.2 Prototype-Based Augmentation

The primary cause of catastrophic forgetting in rPPG DIL is the variation of domain factors such as illumination condition and subject's head motion. In this work, to mitigate catastrophic forgetting, we propose to extract domain prototypes for each task and reproduce domain factors of old tasks during the training on new tasks.

Firstly, we extract the style features from training samples after training on the current task. These style features encompass crucial information of the input videos, such as illumination and the subject's skin color. Existing approaches commonly employ the channel-wise mean and variance of the feature extracted by the shallow layers of the network to represent the style distribution of the input [21, 22]. In accordance with these methods, let $\boldsymbol{h}_{old} \in \mathbb{R}^{C\times T_1\times H_1\times W_1}$ denotes the shallow feature extracted by the backbone from the input $\boldsymbol{x}_{old}$. The style feature $\{\boldsymbol{\mu}_{old}, \boldsymbol{\sigma}_{old}\}$ corresponding to $\boldsymbol{h}_{old}$ can be calculated as follows:

$$\begin{aligned}\boldsymbol{\mu}_{old} &= \frac{1}{T_1H_1W_1}\sum_{t=1}^{T_1}\sum_{h=1}^{H_1}\sum_{w=1}^{W_1}\boldsymbol{h}_{old}^{t,h,w}\\ \boldsymbol{\sigma}_{old} &= \sqrt{\frac{1}{T_1H_1W_1}\sum_{t=1}^{T_1}\sum_{h=1}^{H_1}\sum_{w=1}^{W_1}\left(\boldsymbol{h}_{old}^{t,h,w}-\boldsymbol{\mu}_{old}\right)^2}\end{aligned} \tag{3}$$

Besides rPPG features, the feature extractor also extracts noises introduced by head motion, image compression, *etc.*, which can not be captured by the style features. This noise may be time-varying and erroneously interpreted as rPPG features by the model. Without corresponding labels, how to disentangle them from rPPG features has always been a tough challenge in rPPG measurement. Existing methods mainly rely on adversarial learning [32] or cross-validation [38] for noise modeling, which often introduce additional modules and incur large

amount of computational and memory cost. To tackle this challenge more efficiently, we argue that high-level features extracted by a well-performing model already possess a relatively high SNR, while the aforementioned noise accounts for only a small portion of the energy in the feature maps. Therefore, taking inspiration from traditional image denoising approaches [16, 42], we can employ Singular Value Decomposition (SVD) to efficiently extract the noise features from high-level feature maps. Let $\boldsymbol{z}_{old} \in \mathbb{R}^{C_2 \times T_2}$ denotes the high-level feature extracted by the backbone from input $\boldsymbol{x}_{old}$. The high-level semantic noise $\boldsymbol{n}_{old} \in \mathbb{R}^{C_2 \times T_2}$ corresponding to $\boldsymbol{z}_{old}$ can be extracted as follows:

$$\begin{aligned} \boldsymbol{z}_{old} &= \mathbf{U}_{old}\boldsymbol{\Sigma}_{old}\mathbf{V}_{old}^T \\ \mathbf{M} &= \mathrm{diag}(0, 0, \dots 0, 1, 1, \dots, 1) \\ \boldsymbol{n}_{old} &= \mathbf{U}_{old}(\boldsymbol{\Sigma}_{old} \odot \mathbf{M})\mathbf{V}_{old}^T \end{aligned} \tag{4}$$

where $\odot$ is Hadamard product and $\mathbf{M} \in \mathbb{R}^{C_2 \times C_2}$ is the diagonal mask for noise extraction with first α diagonal elements set to zero. Finally, considering that there may be multiple domains within a single rPPG dataset as discussed in [33], we apply KMeans to group the extracted style and noise features respectively, and obtain K style prototypes $\{\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\}_{k=1}^K$ as well as K noise prototypes $\{\boldsymbol{n}_k\}_{k=1}^K$ for the current task. Different from class prototypes, both of the style and noise features are highly irrelevant to the heart rate, making our domain prototypes more applicable for rPPG measurement where the labels are continuous.

With the domain prototypes of old tasks, we can replay domain factors that only occur in old tasks when training the model on new tasks. Specifically, to reproduce old styles, we employ AdaIN [22] to the shallow feature $\boldsymbol{h}_{new}$ of the new training sample:

$$\boldsymbol{h}_{new}^{style} = \boldsymbol{\sigma}_k \frac{\boldsymbol{h}_{new} - \boldsymbol{\mu}_{new}}{\boldsymbol{\sigma}_{new}} + \boldsymbol{\mu}_k \tag{5}$$

where $\boldsymbol{\mu}_{new}$ and $\boldsymbol{\sigma}_{new}$ are the channel-wise mean and variance of $\boldsymbol{h}_{new}$ as calculated by Eq. (3). $\{\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\}$ is a randomly selected style prototype. For $\boldsymbol{h}_{new}$, this style transfer operation is applied with a probability of p. As for the reproduction of information in the noise prototypes, we can apply a similar operation described in Eq. (4) to mix up the randomly selected noise prototype $\boldsymbol{n}_k$ and the training sample $\boldsymbol{x}_{new}$ with a probability of p:

$$\begin{aligned} \boldsymbol{z}_{new} &= \mathbf{U}_{new}\boldsymbol{\Sigma}_{new}\mathbf{V}_{new}^T \\ \boldsymbol{z}_{new}^{noise} &= \mathbf{U}_{new}(\boldsymbol{\Sigma}_{new} \odot \mathbf{M}')\mathbf{V}_{new}^T + \boldsymbol{n}_k \end{aligned} \tag{6}$$

where $\boldsymbol{z}_{new}$ is high-level feature of $\boldsymbol{x}_{new}$ and $\mathbf{M}' = \mathrm{diag}(1, 1, \dots, 1, 0, 0, \dots, 0)$. Note that these two augmentations are performed independently and randomly, which not only serves to reinforce previous knowledge, but also ensures that the model can acquire new knowledge.

The prototype-based augmentation leverages domain prototypes extracted from the previous tasks to perform feature-level augmentation on training samples of new tasks. This strategy effectively consolidates the past knowledge by

generating pseudo-old samples with domain factors of previous tasks. Meanwhile, these pseudo-old samples also serve as unseen samples for the model, enabling it to learn from a broader range of scenarios. Consequently, this strategy can also enhance the model's generalizability, which is crucial for improving performance of the model in DIL according to [7].

4.3 Prototype-Based Inference Simplification

Similar to humans, models inevitably experience knowledge forgetting in the context of incremental learning. On the other hand, they also become more proficient in specific knowledge of certain tasks. Inspired by the fact that humans usually tackle difficult tasks by transforming them into familiar ones, we can utilize domain prototypes extracted during training to transform test samples into a form that the model is familiar with and can easily process. Specifically, we keep track of the training mean absolute error (MAE) associated with each style prototype used for augmenting the training samples. The style prototype with low training MAE is considered as the style that the model can easily process. During the inference stage, we leverage AdaIN to transfer the test sample $\boldsymbol{x}_t$ into the style that is most familiar to the model:

$$\boldsymbol{h}_t^{style} = \boldsymbol{\sigma}_l \frac{\boldsymbol{h}_t - \boldsymbol{\mu}_t}{\boldsymbol{\sigma}_t} + \boldsymbol{\mu}_l \tag{7}$$

where $\boldsymbol{h}_t$ is the shallow feature of $\boldsymbol{x}_t$; $\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t$ are the channel-wise mean and variance of $\boldsymbol{h}_t$ and $\{\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l\}$ is the style prototype with lowest training MAE.

Meanwhile, we observe that to fight against the augmentation presented in Eq. (6), the model should extract more rPPG information for the feature components with large singular values. In other words, the noise augmentation can further improve the SNR of the high-level features and enhance the model's robustness to these noises. Consequently, there is no need to perform additional noise transfer during the inference stage.

5 Experiments

5.1 Datasets and Evaluation Protocols

Datasets. We use five datasets with various domain factors (lighting condition, head motion, *etc.*) to establish the rPPG DIL setting:

VIPL-HR [36, 37] contains 2378 RGB and 752 NIR videos recorded by 4 cameras across 9 different scenarios. In our experiment, we only use the RGB videos since the RGB modality is more commonly accessible. It is important to note that the frame rates of the videos vary, which can introduce additional distribution shifts.

PURE [49] consists of 60 videos captured by an eco274CVGE camera with six different activities, namely sitting still, talking and four types of head translation and rotation.

UBFC-rPPG [6] comprises 42 uncompressed videos captured using a Logitech C920 HD Pro camera. These videos exhibit variations in lighting conditions, including varying amounts of sunlight and indoor illumination.

BUAA-MIHR [55] contains 165 videos recorded under varying illuminations ranging from 1.0 to 100.0 lux captured by a Logitech HD pro webcam C930E color camera. We only use data with illumination greater than or equal to 6.3 lux because underexposed videos require special algorithms that are not considered in this work.

MMPD [50] comprises 660 mobile phone videos featuring subjects with Fitzpatrick skin types 3-6. The dataset encompasses four distinct lighting conditions, namely LED-high, LED-low, incandescent, and natural lighting. Additionally, it covers four different activities: stationary, head rotation, talking, and walking.

Evaluation Protocols. Firstly, we would like to emphasize that compared with the other 4 folds, samples in VIPL Fold5 are found to have larger head motion and be more challenging, making it a bit more challenging for the model trained on Fold1-4 to generalize well on Fold5. Consequently, we treat Fold5 as a single task while regarding Fold1-4 and the other 4 datasets as 5 separate tasks. This results in a total of 6 tasks derived from the mentioned 5 datasets and we randomly split each task according to the subject-exclusive protocol. Meanwhile, since large-scale data is required to train a robust base model, we use VIPL Fold1-4 task as the initial task.

As for evaluation metrics, the standard deviation of the error (Std), MAE, root mean square error (RMSE) and Pearson's correlation coefficient (R) are adopted to evaluate the performance of video-level HR estimation. The unit for Std, MAE and RMSE is beats per minute (bpm). Furthermore, to evaluate the overall performance of a model under DIL setting, we report the commonly used final incremental performance P_N which can be calculated as follows:

$$P_N = \frac{1}{N}\sum_{j=1}^{N} p_{N,j} \tag{8}$$

where $p_{N,j}$ is a specific metric among Std, MAE, RMSE and R evaluated on the test set of the j-th task after learning all the N tasks. We ran all the incremental experiments 3 times with different task orders and report the mean and standard deviation of these 3 runs.

5.2 Implementation Details

Our proposed method is implemented using PyTorch. Following [59], we first utilize the MTCNN face detector [60] to crop the face region in the first frame and fix the region through the following frames. Subsequently, we sample a certain video into clips with a time window of 160 with step 80 and resize them into 96×96 pixels. For the ground truth rPPG signals, we employ cubic spline interpolation to align them with the corresponding videos.

Table 1: HR estimation results of the backbone effectiveness study. We evaluate our backbone on the VIPL Fold1-4 and MMPD tasks. "EfficientNorm" denotes the normalization module in EfficientPhys [31]. The best results are in bold, and the second best results are underlined.

	VIPL Fold1-4				MMPD			
Methods	Std↓	MAE↓	RMSE↓	R↑	Std↓	MAE↓	RMSE↓	R↑
ViT [12]	15.29	12.52	16.18	0.08	15.94	14.33	18.72	0.12
PhysFormer [59]	8.09	6.23	8.62	0.78	12.14	8.23	12.16	0.67
Uniformer [28]	7.94	5.56	8.19	0.79	<u>9.35</u>	<u>6.05</u>	<u>9.45</u>	<u>0.79</u>
Uniformer + EfficientNorm	<u>7.85</u>	<u>5.18</u>	<u>7.85</u>	<u>0.80</u>	10.91	7.15	10.92	0.71
Uniformer + DiffNorm (ours)	**7.25**	**4.84**	**7.34**	**0.83**	**8.67**	**5.70**	**8.72**	**0.82**

We utilize Uniformer-S [28] as the base model and design a regression head consist of transposed convolutional and convolutional layers. The style and noise features are extracted from the output of the second and fourth stages of Uniformer, respectively. The bottleneck ratio of adapter is set to 0.25.

To train our model, we employ the loss functions introduced in [59] and utilize the Adam optimizer [24] with an initial learning rate of 1e-4 and weight decay of 5e-5. The number of training epoch is 20 on the initial task and 10 on the following tasks. The batch size is 8 for all tasks. Hyperparameters K, p, α are set to 8, 0.5 and 9 respectively.[2] Following [39, 48], random horizontal flipping, spatially resized crop, temporal resampling and image intensity noise are used for data augmentation.

5.3 Effectiveness of Backbone

To validate the effectiveness of our backbone proposed in Sec. 4.1, we compare our model to three baselines: vanilla ViT [12], PhysFormer [59] and vanilla Uniformer [28]. It is worth mentioning that we primarily focus on end-to-end transformer-based models in this work for two reasons. Firstly, non-end-to-end rPPG methods typically involve complex preprocessing procedure to generate the input STMaps, whereas end-to-end methods offer greater accessibility in real-world scenarios. Secondly, the paradigm of a frozen ViT with PEFT modules has demonstrated outstanding performance in continual learning for image classification [47, 53].

To evaluate these models for HR estimation, we conduct intra-task evaluations on two large-scale tasks: VIPL Fold1-4 and MMPD and the results are presented in Tab. 1. We can see that vanilla ViT completely fails to capture rPPG features for both tasks, resulting in poor performance with MAEs exceeding 10 bpm. PhysFormer, a model built upon ViT and specifically designed for rPPG measurement, introduces 3D temporal difference convolution and significantly improves the performance, achieving an MAE of 6.23 bpm on VIPL Fold1-4.

[2] For the influence of the selection of these hyperparameters, please refer to the supplementary material for details.

The third baseline, vanilla Uniformer, further surpasses PhysFormer and get an MAE of 5.56 bpm on VIPL Fold1-4. Building upon Uniformer, we additionally design a DiffNorm module at the input layer. Despite its simplicity, this module proves to be highly beneficial, leading to a substantial reduction in Std, MAE and RMSE, as well as an improvement in R. Similar results can also be observed on MMPD. We further compare our DiffNorm module to the normalization module of EfficientPhys [31] which only extracts dynamic features from input videos. It can be observed that the normalization module in EfficientPhys is less effective on the VIPL Fold1-4 task and even detrimental to the backbone on the more challenging MMPD task due to the absence of appearance features.

To summarize, we find that vanilla self-attention mechanism is insufficient for effectively extracting rPPG features. We believe that this is because local features play a crucial role in capturing the subtle color changes caused by heartbeats. In contrast, PhysFormer and Uniformer additionally introduces 3D convolution to extract these local features. Moreover, the success of PhysFormer and our DiffNorm emphasizes the importance of difference operation and dynamic features. We use Uniformer with a DiffNorm module as the backbone in the subsequent experiments.

5.4 Benchmark Results

We evaluate our method (ADDP) in the rPPG DIL scenario, comparing it with 5 baselines that do not store previous samples: EWC [25], LwF [29], ANCL [23], LAE [14] and S-Prompts [52]. We replace the prompt pool with an adapter pool when employing S-Prompts to Uniformer. This is reasonable as prompts can be transformed into a similar form as adapters according to [17]. Similar to frequency cross entropy loss [38], power spectral density (PSD) of the predicted rPPG signal is considered as the predicted logit for LwF and LAE. Additionally, the joint training that learns all tasks together and naive full finetuning the model without any countermeasure to forgetting are recognized as the upper and lower bounds of the performance in our rPPG DIL experiments.

The HR estimation results are presented in Tab. 2. Compared with the lower bound, both EWC and LwF effectively alleviate forgetting and achieve better last incremental performance. However, the strict stability constraints imposed on the model lead to a reduction in plasticity and forward transfer capability. ANCL further introduces an auxiliary model for plasticity regularization and slightly improves the performance. LAE is a state-of-the-art approach for class incremental learning. However, all of its three strategies (online adapter, offline adapter and experts ensemble) fail to achieve satisfactory performance in our rPPG DIL. We attribute this phenomenon to two inherent disadvantages of LAE. Firstly, the stability of LAE is highly dependent on the value of weight decay for updating the offline adapter. Secondly, the logit for ensemble prediction is hard to define in rPPG regression problem.

S-Prompts achieves the second lowest Std_N, MAE_N, RMSE_N and the second highest R_N among these baselines. Nevertheless, it still lags behind our method and the upper bound by a considerable margin. This is because S-Prompts learns

Table 2: HR estimation results on the rPPG DIL protocol. Std_N represents the final incremental Std and the same applies to MAE_N, RMSE_N as well as R_N.

Methods	$\text{Std}_N \downarrow$	$\text{MAE}_N \downarrow$	$\text{RMSE}_N \downarrow$	$\text{R}_N \uparrow$
Upper Bound	5.25±0.09	3.45±0.09	5.31±0.24	0.85±0.01
Lower Bound	8.73±0.79	6.54±0.27	9.46±0.49	0.69±0.02
EWC [25]	6.88±0.89	4.51±0.33	7.02±0.86	0.76±0.04
ANCL-EWC [23]	6.87±0.36	4.39±0.08	6.97±0.36	0.77±0.02
LwF [29]	6.59±0.14	4.38±0.10	6.30±0.67	0.78±0.02
ANCL-LwF [23]	6.37±0.28	4.13±0.13	6.12±0.25	0.80±0.01
LAE-online [14]	6.97±0.56	4.70±0.38	7.13±0.56	0.78±0.03
LAE-offline [14]	12.06±0.86	7.94±0.35	12.36±0.96	0.59±0.04
LAE-ensemble [14]	12.37±0.42	8.07±0.24	12.61±0.51	0.55±0.01
S-Prompts [52]	<u>5.77</u>±0.18	<u>3.85</u>±0.05	<u>5.82</u>±0.19	<u>0.81</u>±0.01
ADDP	**5.56**±0.04	**3.70**±0.07	**5.59**±0.02	**0.83**±0.01

Table 3: HR estimation results of the ablation study on rPPG DIL protocol. "Style.", "Noise." and "Sim." are style, noise prototype-based augmentation and prototype-based inference simplification respectively. "TS" means we replace our inference simplification with test-time style shifting proposed by [40].

Style.	Noise.	Sim.	$\text{Std}_N \downarrow$	$\text{MAE}_N \downarrow$	$\text{RMSE}_N \downarrow$	$\text{R}_N \uparrow$
×	×	×	7.06±0.90	4.59±0.53	7.24±0.96	0.78±0.03
√	×	×	6.68±0.67	4.23±0.21	6.83±0.60	0.73±0.06
√	√	×	6.66±0.61	4.10±0.20	6.61±0.63	0.75±0.05
√	√	TS	<u>5.71</u>±0.01	<u>3.91</u>±0.04	<u>5.89</u>±0.03	<u>0.81</u>±0.00
√	√	√	**5.56**±0.04	**3.70**±0.07	**5.59**±0.02	**0.83**±0.01

each task independently, lacking knowledge transfer capability and resulting in relatively poor performance on tasks that do not correspond to the selected prompts (adapters in this work). Additionally, S-Prompts utilizes KMeans to cluster features with different labels into domain centroids, which leads to feature interference and undermines the accuracy of task prediction. By contrast, ADDP finetunes a single group of adapters to facilitate knowledge sharing and avoid task prediction, while domain prototypes further aid in consolidating past knowledge and simplifying the inference. Therefore, our method finally achieves an Std_N of 5.56 bpm, an MAE_N of 3.70 bpm, an RMSE_N of 5.59 bpm and an R_N of 0.83, significantly narrowing the performance gap with the upper bound.

5.5 Ablation Study

The proposed ADDP consists of three main designs: style and noise prototype-based augmentation as well as inference simplification. We conduct ablation studies on these components and present the results in Tab. 3. The first row is the baseline model that simply utilizes adapter to finetune the backbone.

We observe that the performance improves when style and noise prototype-based augmentation are employed. This demonstrates that both prototypes ef-

fectively remind the model of the domain factors of previous tasks, thereby mitigating forgetting. It is worth mentioning that we could also directly employ centroids of raw features to replay domain factors. However, this would results in feature interference among samples with different heart rates. In contrast, our style and noise features are almost irrelevant to the heart rate, allowing us to employ KMeans to group them without introducing feature interference.

We also see a significant improvement in performance when utilizing inference simplification. This makes sense because this strategy enables the transfer of potentially forgotten styles into a more familiar one for the model. We have noticed that [40] also proposes a test-time style shifting (TS) strategy to address domain generalization (DG) by shifting the style of test samples to the nearest source domain. To compare our inference simplification strategy with TS, we replaced it with TS in our experiments. The results in the fourth row of Tab. 3 show that TS brings less improvement compared to our strategy. We attribute it to the difference between DG and DIL. In the case of DIL, the model has already seen almost all styles of test samples, rendering the nearest style shifting less effective during inference. Furthermore, the model may have already forgotten the nearest style, resulting in a negative impact on some tasks when employing style shifting. In contrast, we utilize the style prototype with the lowest training MAE to ensure that the model can effectively process the transferred samples.

6 Conclusion

This paper focuses on the domain incremental learning for rPPG measurement (rPPG DIL), which has never been explored before. To tackle this challenge, we present a practical rehearsal-free method named ADDP. Adapter finetuning is employed for efficiently adapting the model to new tasks while keeping the stability of the model. Besides, we design a simple yet effective difference normalization (DiffNorm) module to integrates the appearance and dynamic features for the backbone. To mitigate forgetting, we design rPPG-friendly domain prototypes and propose prototype-based augmentation, which generates pseudo-old samples with domain factors of previous tasks. Furthermore, we employ an inference simplification strategy to transform challenging tasks into more manageable ones. To compare our method and existing continual learning algorithms, we establish the first rPPG DIL protocol. Extensive experiments demonstrate that ADDP achieves satisfactory performance superior to other baselines.

Limitations: Firstly, ADDP relies on a well-performing backbone, which may be unavailable when the initial task is not large and diverse enough. In such cases, more tuning on the backbone may be necessary during subsequent learning. Secondly, our method has not been evaluated in the online continual learning scenario, a more realistic setting where the model learns from a single-pass data stream and can only access each batch of data once. Due to the non-stationary nature of the stream, this setting is also more challenging and may require additional designs.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62172381).

References

1. Ahn, H., Kwak, J., Lim, S., Bang, H., Kim, H., Moon, T.: Ss-il: Separated softmax for incremental learning. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 844–853 (2021)
2. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget. In: Proceedings of the European conference on computer vision (ECCV). pp. 139–154 (2018)
3. Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient based sample selection for online continual learning. Advances in neural information processing systems **32** (2019)
4. Bang, J., Kim, H., Yoo, Y., Ha, J.W., Choi, J.: Rainbow memory: Continual learning with a memory of diverse samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8218–8227 (2021)
5. Benezeth, Y., Li, P., Macwan, R., Nakamura, K., Gomez, R., Yang, F.: Remote heart rate variability for emotional state monitoring. In: 2018 IEEE EMBS international conference on biomedical & health informatics (BHI). pp. 153–156. IEEE (2018)
6. Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., Dubois, J.: Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognition Letters **124**, 82–90 (2019)
7. Cai, R., Cui, Y., Li, Z., Yu, Z., Li, H., Hu, Y., Kot, A.: Rehearsal-free domain continual face anti-spoofing: Generalize more and forget less. arXiv preprint arXiv:2303.09914 (2023)
8. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. In: Proceedings of the European conference on computer vision (ECCV). pp. 233–248 (2018)
9. Chen, W., McDuff, D.: Deepphys: Video-based physiological measurement using convolutional attention networks. In: Proceedings of the european conference on computer vision (ECCV). pp. 349–365 (2018)
10. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. IEEE Transactions on Biomedical Engineering **60**(10), 2878–2886 (2013)
11. De Haan, G., Van Leest, A.: Improved motion robustness of remote-ppg by using the blood volume pulse signature. Physiological measurement **35**(9), 1913 (2014)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Du, J., Liu, S.Q., Zhang, B., Yuen, P.C.: Dual-bridging with adversarial noise generation for domain adaptive rppg estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10355–10364 (2023)
14. Gao, Q., Zhao, C., Sun, Y., Xi, T., Zhang, G., Ghanem, B., Zhang, J.: A unified continual learning framework with general parameter-efficient tuning. arXiv preprint arXiv:2303.10070 (2023)

15. Goodfellow, I.J., Mirza, M., Xiao, D., Courville, A., Bengio, Y.: An empirical investigation of catastrophic forgetting in gradient-based neural networks. arXiv preprint arXiv:1312.6211 (2013)
16. Guo, Q., Zhang, C., Zhang, Y., Liu, H.: An efficient svd-based method for image denoising. IEEE transactions on Circuits and Systems for Video Technology **26**(5), 868–880 (2015)
17. He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G.: Towards a unified view of parameter-efficient transfer learning. arXiv preprint arXiv:2110.04366 (2021)
18. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Lifelong learning via progressive distillation and retrospection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 437–452 (2018)
19. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 831–839 (2019)
20. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
21. Huang, W., Chen, C., Li, Y., Li, J., Li, C., Song, F., Yan, Y., Xiong, Z.: Style projected clustering for domain generalized semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3061–3071 (2023)
22. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
23. Kim, S., Noci, L., Orvieto, A., Hofmann, T.: Achieving a better stability-plasticity trade-off via auxiliary networks in continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11930–11939 (2023)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
25. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
26. Lee, S., Ha, J., Zhang, D., Kim, G.: A neural dirichlet process mixture model for task-free continual learning. arXiv preprint arXiv:2001.00689 (2020)
27. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)
28. Li, K., Wang, Y., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unified transformer for efficient spatiotemporal representation learning. arXiv preprint arXiv:2201.04676 (2022)
29. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
30. Liu, X., Fromm, J., Patel, S., McDuff, D.: Multi-task temporal shift attention networks for on-device contactless vitals measurement. Advances in Neural Information Processing Systems **33**, 19400–19411 (2020)
31. Liu, X., Hill, B., Jiang, Z., Patel, S., McDuff, D.: Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5008–5017 (2023)

32. Lu, H., Han, H., Zhou, S.K.: Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12404–12413 (2021)
33. Lu, H., Yu, Z., Niu, X., Chen, Y.C.: Neuron structure modeling for generalizable remote physiological measurement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18589–18599 (2023)
34. Malepathirana, T., Senanayake, D., Halgamuge, S.: Napa-vq: Neighborhood-aware prototype augmentation with vector quantization for continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11674–11684 (2023)
35. McDuff, D.: Camera measurement of physiological vital signs. ACM Computing Surveys **55**(9), 1–40 (2023)
36. Niu, X., Han, H., Shan, S., Chen, X.: Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14. pp. 562–576. Springer (2019)
37. Niu, X., Shan, S., Han, H., Chen, X.: Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. IEEE Transactions on Image Processing **29**, 2409–2423 (2019)
38. Niu, X., Yu, Z., Han, H., Li, X., Shan, S., Zhao, G.: 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. pp. 295–310. Springer (2020)
39. Niu, X., Zhao, X., Han, H., Das, A., Dantcheva, A., Shan, S., Chen, X.: Robust remote heart rate estimation from face utilizing spatial-temporal attention. In: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019). pp. 1–8. IEEE (2019)
40. Park, J., Han, D.J., Kim, S., Moon, J.: Test-time style shifting: Handling arbitrary styles in domain generalization. arXiv preprint arXiv:2306.04911 (2023)
41. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Optics express **18**(10), 10762–10774 (2010)
42. Rajwade, A., Rangarajan, A., Banerjee, A.: Image denoising using the higher order singular value decomposition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(4), 849–862 (2012)
43. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
44. Robins, A.: Catastrophic forgetting, rehearsal and pseudorehearsal. Connection Science **7**(2), 123–146 (1995)
45. Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks. arXiv preprint arXiv:1606.04671 (2016)
46. Serra, J., Suris, D., Miron, M., Karatzoglou, A.: Overcoming catastrophic forgetting with hard attention to the task. In: International conference on machine learning. pp. 4548–4557. PMLR (2018)
47. Smith, J.S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., Kira, Z.: Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11909–11919 (2023)

48. Speth, J., Vance, N., Flynn, P., Czajka, A.: Non-contrastive unsupervised learning of physiological signals from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14464–14474 (2023)
49. Stricker, R., Müller, S., Gross, H.M.: Non-contact video-based pulse rate measurement on a mobile service robot. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication. pp. 1056–1062. IEEE (2014)
50. Tang, J., Chen, K., Wang, Y., Shi, Y., Patel, S., McDuff, D., Liu, X.: Mmpd: Multi-domain mobile video physiology dataset. arXiv preprint arXiv:2302.03840 (2023)
51. Wang, W., Stuijk, S., De Haan, G.: Exploiting spatial redundancy of image sensor for motion robust rppg. IEEE transactions on Biomedical Engineering **62**(2), 415–425 (2014)
52. Wang, Y., Huang, Z., Hong, X.: S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. Advances in Neural Information Processing Systems **35**, 5682–5695 (2022)
53. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: European Conference on Computer Vision. pp. 631–648. Springer (2022)
54. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 139–149 (2022)
55. Xi, L., Chen, W., Zhao, C., Wu, X., Wang, J.: Image enhancement for remote photoplethysmography in a low-light environment. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). pp. 1–7. IEEE (2020)
56. Xue, M., Zhang, H., Song, J., Song, M.: Meta-attention for vit-backed continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 150–159 (2022)
57. Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. arXiv preprint arXiv:1905.02419 (2019)
58. Yu, Z., Li, X., Zhao, G.: Facial-video-based physiological signal measurement: Recent advances and affective applications. IEEE Signal Processing Magazine **38**(6), 50–58 (2021)
59. Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P.H., Zhao, G.: Physformer: Facial video-based physiological measurement with temporal difference transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4186–4196 (2022)
60. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters **23**(10), 1499–1503 (2016)
61. Zhu, F., Cheng, Z., Zhang, X.y., Liu, C.l.: Class-incremental learning via dual augmentation. Advances in Neural Information Processing Systems **34**, 14306–14318 (2021)
62. Zhu, F., Zhang, X.Y., Wang, C., Yin, F., Liu, C.L.: Prototype augmentation and self-supervision for incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5871–5880 (2021)