参赛学生姓名：林聿扬

中学：杭州钱江贝赛斯国际学校

省份：浙江

国家/地区：中国

指导老师姓名：王胤，邓水光

指导老师单位：浙江大学

论文题目：Class-Imbalanced Semi-Supervised Learning for Robust rPPG-based Heart Rate Estimation

Class-Imbalanced Semi-Supervised Learning for Robust rPPG-based Heart Rate Estimation

Yuyang Lin
BASIS International School Hangzhou
Hangzhou, Zhejiang, China
yuyang.lin14883-bihz@basischina.com

Abstract

Remote Photoplethysmography (rPPG) has been crucial physiological measurements to be implemented both in clinical and daily scenarios due to its convenience and non-contact characteristic. However, the nature of rPPG makes it hard to collect measurement data, resulting in both scarce and imbalanced data across many datasets. This led to existing models learning patterns biased towards common heart rates, resulting in inferior performance on elevated or abnormal cases: the long-tail cases. In light of this, we propose a class-imbalanced semi-supervised learning approach integrating the CoSSL paradigm together with domain-specific explicit rPPG priors. While the use of CoSSL inevitably pivots the task into a classification task, an additional Label Distribution Smoothing (LDS) adds the reegression-like continuity of rPPG back into classification, resulting in superior model performances. Experiments on VIPL-HR, UTKFace, and Yelp Review datasets demonstrate that our method consistently outperforms state-of-the-art baselines, validating its effectiveness in addressing both imbalanced distributions and generalization to diverse scenarios.

Index term: rPPG, Class-imbalanced semi-supervised learning, long-tail

1. Introduction

Physiological measurements, such as heart rate (HR), are crucial indicators of human emotional state and cardiovascular conditions. Monitoring HR aids in both medical conditions and daily scenarios. ECG is widely used for accurate physiological measurements; however, as a contact-based method, it is inconvenient and may cause skin irritation or other medical conditions with long-term usage [14]. Remote Photoplethysmography (rPPG), on the other hand, analyzes the subtle changes in skin color to estimate blood volume variations and subsequent physiological activities.

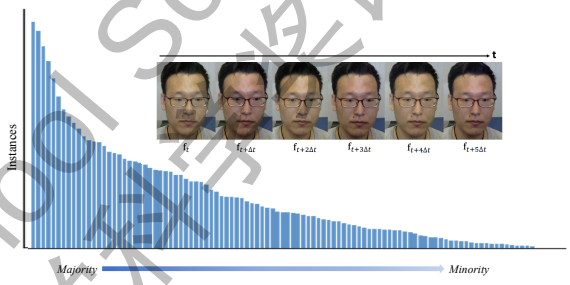

Figure 1. VIPL [31] dataset ranked on the number of data samples of each HR reading. The heavily right-skewed distribution indicates a imbalance in data samples among classes, posing the long-tail problem. In addition, an example of frames from one VIPL video clip is shown for reference.

This method provides a non-contact solution for physiological measurements. Compared to contact photoplethysmography (cPPG) and ECG, rPPG grants a greater range of applicable scenarios without the necessary constraint of a contacting device.

rPPG uses camera to remotely track the subtle variations in skin color from the reflection of light beneath the skin, indicating physiological signs. The change in reflection wavelength ultimately indicates the blood volume pulse in the region of interest. Existing studies mostly use camera to film a short clip of subjets' face in order to retrieve such data. The video clips are then processed and analyzed through either conventional signal processing or deep learning approaches to study the pattern presented within. While some studies focused on signals in the infrared or near-infrared spectrum [9, 29, 34, 48], most have focused on natural or ambient lighting scenarios, using commercial grade webcams or phone cameras. The study of rPPG using easily accessible devices in normal life further adds to its potential to be implemented in real-world applications.

As such, rPPG studies often focus on model abilities to generalize to more diverse, out-of-lab, and unseen scenarios [28, 40, 53, 62], increasing model robustness through tackling priors relating to lighting, movements, skin conditions, and so on. However, existing studies have overlooked the prevalence of the long-tail problem in rPPG datasets.

Data collection has long been a critical issue for rPPG-related research. A proper data point requires subjects to both film their face and wear a contact device for ground-truth measurements. This raises the difficulty of collecting as well as ethical concerns behind it, which results in many datasets having no more than 50 subjects. In addition to the small datasets, the long-tail problem refers to the lack of enough samples representing large ranges of HR values on the two sides of the data distribution in current datasets. The lack of elevated or abnormal HR that are lower than normal in most datasets [1] leads to an imbalanced distribution (Fig. 1) of data points overly centered in the settled HR range. According to [8], model performances may be severely hampered due to the lack of proper representations of HR that appear on the two tails of the distribution. This may decrease model robustness when encountering unseen variations due to disease or exercise – scenarios that lead to higher or lower than average HR readings. As a result, resolving the long-tail problem in rPPG is one of the crucial stepping stones in which to achieve robust generalization in real-world applications.

In light of this. We address the long-tail problem through the use of an imbalanced semi-supervised learning (SSL) paradigm. In sum, we implement the CoSSL paradigm [15] for imbalanced SSL while also maintaining the rPPG priors necessary to strengthen task-specific model ability. To fully utilize the CoSSL paradigm, we treat rPPG as a classification task with each integer HR value as a class. In CoSSL, Fan et al. implement a decoupled training process, separating representation learning and classifier learning without gradient exchange. In doing so, the training of the two sections is completely independent of each other. For representation learning, decoupling helps the feature encoder learn a class-agnostic feature space that aims to better capture the overall data pattern; for classifier learning, decoupling prevents the classifier head from learning data biased to the popular class. The co-learning framework provided here boosts the model's ability in learning features independent from the class imbalance, proven empirically to be successful [19]. We also integrate the Tail-class Feature Enhancement (TFE) module from CoSSL to further support the classifier learning process. Specifically, TFE provides a greater diversity of tail-classes' samples through augmenting unlabeled data which the feature encoder deems similar to the feature of that tail class in interest. As a result, TFE emulates a class-balanced data distribution for classifier training that prevents the influence of an imbalanced data distribution. In addition, we used Label Distribution Smoothing (LDS) [55] to regulate the probability of applying TFE on every batch of data, so that we fully exploit the continuity nature of the rPPG task as a regression. LDS applies a convolution on the original data distribution to portray a more reasonable representation of how each class is weighted in a dataset, adjusting to a more accurate probability for the application of TFE during classifier training.

In addition, we also utilize necessary rPPG priors for the CoSSL paradigm to full adapt to the rPPG task specifically. According to [39, 62], prior knowledge of physiological signals independent of frame rates, brightness, motion, time sequences, etc. is implemented during data augmentation and image processing. Taken as important preset principles, these priors are used to further increase the model's generalization ability to different apparatus domains, increasing the feature encoder's ability to withstand different noises and influences.

Our contributions are threefold:

- We highlight the long-tail distribution problem in rPPG datasets, showing its critical impact on model generalization to abnormal or elevated HRs.
- We propose a class-imbalanced semi-supervised learning framework tailored for rPPG, integrating the CoSSL paradigm with domain-specific rPPG priors. This design effectively decouples representation learning and classifier learning, while enhancing tail-class representation through TFE.
- We apply Label Distribution Smoothing (LDS) to restore the regression-like continuity of HR estimation within the classification setting, leading to superior robustness and performance across diverse benchmarks (VIPL-HR, UTKFace, Yelp Review).

2. Related Works

2.1. Remote Physiological Measurements

Traditional rPPG methods manually observe physiological patterns in regard to the reflection of skin color. GREEN [49] finds that the green channel in comparison to red and blue channels creates a higher signal-to-noise ratio (SNR) and thus builds the initial groundwork for the remote PPG task. Blind source separation (BSS) methods such as ICA [35] and PCA [24] are dimensionality reduction techniques implemented on RGB temporal signals for identifying desired physiological signals through noises. CHROM [13] and POS [50] are handcrafted methods that involve the projection of signals on to orthogonal planes in order to adapt to skin tone and illumination variations. As deep learning becomes increasingly prevalent, studies also implement end-to-end neural networks for more robust means of extracting physiological patterns. Early examples include DeepPhys [7], a landmark applica-

tion of CNN on different frames to retrieve variations. Afterward, TS-CAN [25] extends on DeepPhys by implementing an attention module while predicting pulse and respiration jointly. As studies progress, more studies have focused on the ability of models to capture spatial-temporal variations. RhythmNet [31] achieves this by creating a spatial-temporal map (STMap) that uses pixel values as a generalization of regions of interest for every designated span of frames. PhysNet [56] and rPPGNet [57] further enhance on models' ability to capture spatial-temporal representations. BVPNet [12] is focused on the prediction of blood volume pulses (BVP) waveform sequences. In addition to 2D-CNN and 3D-CNN methods, others such as Dual-GAN [27] harness the unique model architecture of generative adversarial network to study and reduce noise within the video. PhysFormer [59] and others [42, 58] use transformer architecture to increase the attention span to capture longer time-ranged dependencies. NEST [28] further uses domain adaptation techniques to increase model robustness in few-shot and zero-shot scenarios.

2.2. Semi-supervised Learning

Given the case where data labeling is particularly expensive or demands professional expertise, while a large field of unlabeled data is easy to get, SSL becomes a potent method to use in order to increase model generalization ability. Existing methods for SSL largely consist of consistency regularization and pseudo-labeling. Consistency regularization refers to the overarching idea that the model should pertain the same output despite receiving inputs subject different augmentation methods in vision tasks. Examples of consistency regularization include [2, 36, 37]. Applying this method allows the model to adapt to different noises that disturb the relevant feature of the task, increasing the model's ability to generalize to unseen cases. Pseudo-labeling utilizes the knowledge of the model trained with scare labeled data to generate fake artificial for unlabeled data in order to bootstrap training data with confident unlabeled data [22]. Proceeding studies combine these two techniques and formulate stronger methods to exploit unlabeled data for model training [3, 4, 11, 41, 60]. [21] uses a temporal ensemble framework to achieve a consistency between epochs. [46] improves on the former by implementing exponential moving average on weights instead of label predictions, further strengthening it at low data scenario. [10] is one of the first to apply contrastive concepts in unlabeled data to be used in deep regression. [17] takes on a similar idea and adapts the deep regression as a ranking classification. However, methods above assume a balanced class distribution; that is, each class has similar number of data, which is not the typical case in real-world applications. [18] identifies such problem as class-imbalanced semi-supervised learning (CISSL) and proposed a novel suppressed consistency loss that reduces the effect of consistency regularization on edge cases with few data. [16] uses an adaptive thresholding for the confidence of pseudo-labels in order to inhibit its effect on minority cases. [20] uses an algorithm called distribution aligning refinery of pseudo-labels to optimize pseudo-labeling process. [23] proposes an additional classifier layer called auxiliary balanced classifier to train the model in a class-balanced manner.

2.3. semi-supervised learning in rPPG

Because of the regression nature of the rPPG task, very limited studies have used SSL paradigm for model training despite the scare labeled data. Consistency Regularization and pseudo-labeling are built on the foundation of innately distinctive categories and labels, making it incompatible to regression. Existing studies that aim to utilize unlabeled data mainly focus on self-supervised learning and contrastive learning techniques that explore rPPG-specific priors [26, 38, 39, 44, 45, 62]. [54] uses curriculum pseudo-labeling as an adaptation to rPPG using SSL; however, few studies have discussed the imbalanced nature of rPPG datasets and how CISSL should be implemented.

3. Methodology

In this study, we combined the CoSSL [15] paradigm, which is the co-learning of representation and classifier in CISSL, with rPPG explicit priors. The method combines the benefit of SSL in using unlabeled data when at a small data size while also retaining crucial task-specific context for applying to the rPPG measurement task. In addition, we adapt the rPPG problem to a classification task in order to adapt to SSL techniques; meanwhile, we implement the label distribution smoothing (LDS) [55] technique to take advantage of the continuity found in regression.

3.1. Co-learning Paradigm

In order for rPPG measurement task to regarded as a CISSL, let $\mathcal{X} = \{(\mathbf{x}_n, y_n);\ n \in (1, \ldots, N)\}$ and $\mathcal{U} = \{(\mathbf{u}_m);\ m \in (1, \ldots, M)\}$ be the labeled and the unlabeled data, respectively, where N_i and M_i denote the number of labeled and unlabeled data points for class i, respectively, in a total of k classes. Here, $(\mathbf{x}_n, \mathbf{u}_n) \in \mathbb{R}^{W \times H \times C}$ are spatial-temporal maps (STMaps) using RGB channel [31, 53], and y_n is the HR ground truth for each $\mathbf{x}_n$ labeled STMap. The end goal of CoSSL is to train a feature encoder $g_f(\cdot)$ and a final classifier $h_{CL}(\cdot)$ independently without sharing gradient parameters at the same time; meanwhile, the encoder $g_f(\cdot)$ and $h_{CL}(\cdot)$ can also be connected together and utilized when creating pseudo-labels. To that end, each module could prevent biased inputs from another while also harnessing one others' strength.

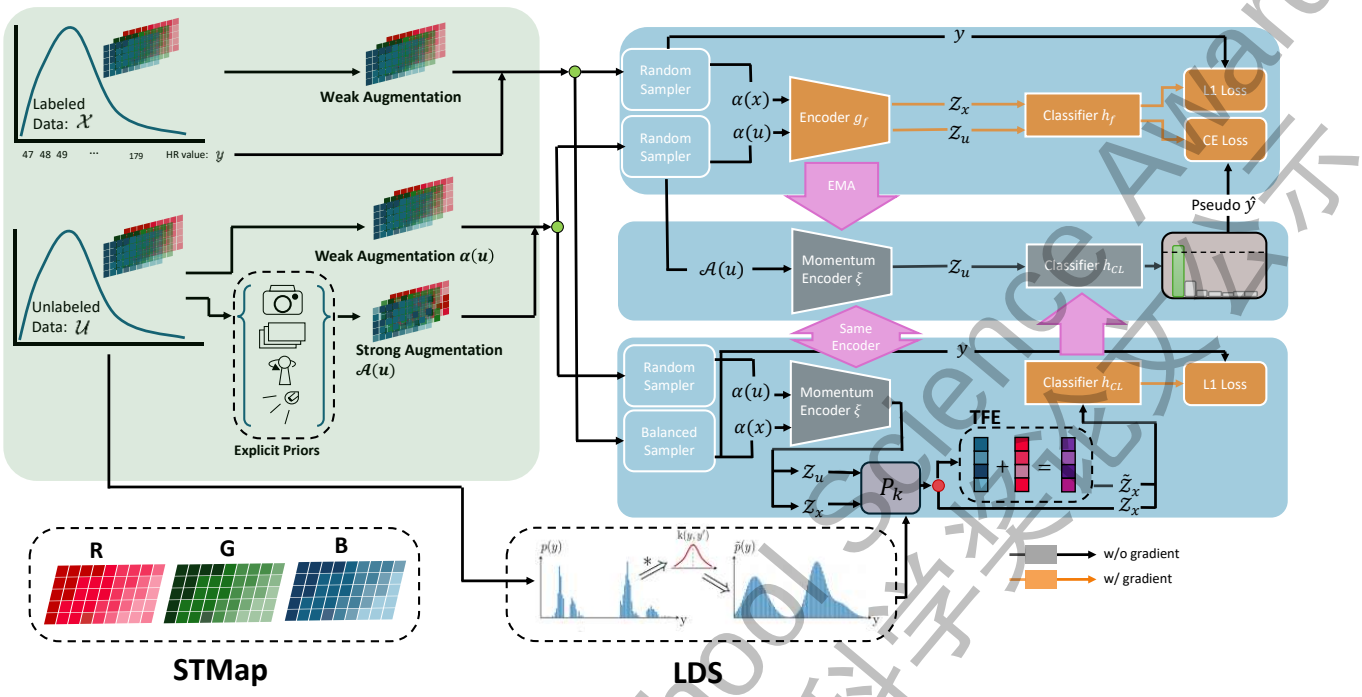

Figure 2. Our methods integrate CoSSL paradigm (in light blue) along with domain-specific explicit priors of rPPG (in light green). Subsequent implementation of LDS module for the probability of P_k pertains the continuity nature of the rPPG task. Part of the pipeline illustration referenced from CoSSL [15]. images for LDS taken from [55].

The training process begins with initializing both the encoder and classifier modules. The encoder module is first initialized through a regular round of SSL training using g_f and a randomly initialized auxiliary classifier h_f. Here, CoSSL allows a flexible room for the specific SSL framework implemented. Examples of viable frameworks include MixMatch [4], ReMixMatch [3], FixMatch [41], FlexMatch [60], and so on. For the sake of consistency and simplicity, we implement FixMatch as our baseline SSL method and perform the initial representation learning. As a result, g_f learns a generalizable feature encoder without the influence of h_{CL} and is thus free from class-imbalanced bias. An exponential moving average (EMA) decay is implemented on the encoder and is passed to classifier module for later use. The classifier module is initialized through a round of Tail-Class Feature Enhancement (TFE), proposed by CoSSL. In order to increase the diversity and the count of data samples for classes that originally contain little existing labeled data, TFE is introduced. It uses logits from other unlabeled samples of a similar pseudo-label class as additional noise to be added onto labeled samples that are known to belong to the class in interest.

$$\tilde{z} = \lambda\xi(\mathbf{x}_l) + (1-\lambda)\xi(\mathbf{u}_j) \tag{1}$$

where $\mathbf{x}_l$ is a labeled STMap in class l, $\mathbf{u}_j$ is an unlabeled STMap predicted to have a high logits for the same class l, λ is a fusion factor created using beta distribution, ξ is the passed down EMA encoder from the encoder module, and $\tilde{z}$ is the resulting newly augmented features based on features of $\mathbf{x}_l$ and $\mathbf{u}_j$, which is $\xi(\mathbf{x}_l)$ and $\xi(\mathbf{u}_j)$ respectively. Such TFE is used to only at a probability value during the training epochs; for initialization, new feature maps are created so that each class contains the same number of feature maps as the class containing most labeled data points. In both cases, TFE helps the classifier training remain a class-balanced training, independent of the influence of a imbalanced class distribution, since TFE bootstraps classes with fewer data samples. After initialization, the model enters the formal training epochs, as shown in Figure 2. The encoder module is first trained using labeled and unlabeled data under FixMatch formality. The EMA encoder is then passed down to the classifier module and trains it along with TFE. The resulting bias-free classifier is then passed together with the EMA encoder to the pseudo-labeling stage, where artificial labels are generated using generalizable features and bias-free classifiers. The pseudo-labels are then passed back to the encoder module as labels for the unlabeled section of the data and complete the whole epoch of training. In this way, the CoSSL paradigm leverages both the benefit of decoupling the two modules and the connectivity between the

two when using to produce pseudo-labels.

Note that due to the large span of possible HR values for physiological measurements, there exist a large number of classes k, typically more than 100 classes for capturing HR values in extreme scenarios, which strongly hinders the pseudo-labeling process in FixMatch. The large amount of classes k makes the supplementary classifier h_f and end product classifier h_{CL} difficult to produce a confident classification with logits of one class reaching the preset threshold τ. To that end, we implement an additional sharpening technique from MixMatch [4] that boost the differences between logits of different classes

$$f_{shar}(p_i, T) = p_i^{\frac{1}{T}} / \sum_{j=1}^{k} p_j^{\frac{1}{T}} \tag{2}$$

where p is the predicted logits for each class, T is a hyperparameter. The resulting $p_{sharpen} = f_{shar}(p, T)$ is then used to determine whether the pseudo-label is confident enough. This significantly help both classifier heads to perform properly in getting pseudo-labels despite the large k.

3.2. Label Distribution Smoothing

While CoSSL paradigm mostly captures the advantage of the model trained as a classification task, we also incorporate the utility of the unique characteristics of regression tasks: continuity. This is based on the assumption that the physiological signal and pattern for a certain HR should be similar to other measurements that are close to such HR value (e.g. a HR measurement of 100 should exhibit similar physiological patterns in comparison to another HR measurement of 101). That is to say, the class of a certain HR should resemble some connection to other classes in a few HRs away, and there should not be hard boundaries distinguishing two adjacent classes. In addition to this, we find that current datasets severely lack data samples with HRs that are either elevated or abnormally low. This, combined with the lack of continuity representation, often leads to unreasonable probabilities when tackling class balancing issues. As previously stated in section 3.1, the TFE takes a class-based probability to be enacted during training. Ideally, the more data samples a class has; the less it requires to perform TFE to sustain data diversity. As such, the CoSSL paradigm defines such probability as

$$P_i = \frac{N_{max} - N_i}{N_{max}} \tag{3}$$

where N_{max} is the largest number of data samples per class among all k classes, N_i is the number of data samples for class i, and P_i is the probability for feature maps belonging to class i to enact TFE. Such method, due to the two drawbacks stated, does not fit perfectly for rPPG-specific task. As a result, we implement the LDS [55] to remedy this. LDS offers a smoothing for the data distribution. This is done through a convolution using a symmetric kernel such as a gaussian or a Laplacian kernel. Here, we use gaussian kernel.

$$\tilde{N}(y') \triangleq \int_{\mathcal{Y}} k(y, y') N(y) dy \tag{4}$$

and

$$P_i = \frac{N_{max} - \tilde{N}(i)}{N_{max}} \tag{5}$$

where $N(y)$ is the count of label y in the train dataset and $\tilde{N}(y')$ is the effective density of label y'. This convolution process subsidizes classes with little data by averaging data counts from few neighboring classes on both sides of the distribution. This process, as later shown in ablation, proves significant.

3.3. Explicit Priors

In addition to that, we assume some explicit priors related to rPPG context for better model generalization ability according to [39, 62]. Given that ST is the raw STMap from [31]. Camera prior assumes that the same subject should give the same HR readings despite measuring on different camera setups. Since videos normally have gamma correction to offset camera illuminance difference to real world illumination, undoing such correction by a random power γ simulates the noise of using different camera.

$$ST_\gamma = (ST)^\gamma, \ \gamma \in [0.8, 2.2] \tag{6}$$

The gamma range is referenced from [6, 62]. The skin and light prior assumes that skin tone and illumination setting may change skin appearance for similar HR values. This noise is emulated by taking dot product of RGB channels with a random matrix.

$$ST_l = \begin{bmatrix} R_l \\ G_l \\ B_l \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{7}$$

where $\{a_{11}, ..., a_{33}\} \in [-0.5, 0.5]$. The motion prior rules out the effect of head movements on physiological patterns. Such noise is arbitrarily added by shuffling the Regions of Interest of a frame, which is the height of the STMap.

$$ST_m = Shuffle(R_1, R_2, ..., R_{25}) \tag{8}$$

where $Shuffle()$ is the random permutation of ROIs. Frame rate prior rules out the effect of different camera frame rates on physiological patterns. This noise is created through down sampling the frames and then cubic interpolating it.

$$ST_f = Cubic(Down(S^{i,j})), i \in \{0, 1, ..., W\}, j \in \{0, 1, ..., C\} \tag{9}$$

where W is the number of ROIs and C is the number of color channels. These priors altogether are implemented in the data augmentation process, where the SSL frameworks such as FixMatch require strongly augmented data for the training to apply consistency regularization along with weakly augmented data. Applying these priors effectively minimizes the loss the model may have due to these four factors.

4. Experiments

4.1. Dataset

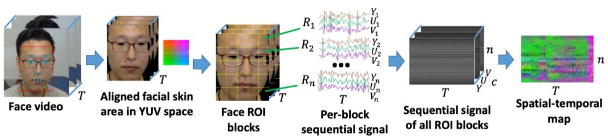

Figure 3. An illustration of spatial-temporal map generation from a face video (adapted from [31]). Faces are aligned and transformed to YUV space, divided into n ROI blocks, and per-channel average color values are computed to form sequential signals concatenated into a spatial-temporal map.

VIPL is a large-scale multi-model HR dataset that has been dedicated to emulate complex situations and less constrained settings through variations in subjects, head movements, illumination status, and camera types [30]. To be specific, VIPL contains a total of 2,379 color videos and 752 near-infrared videos from 107 participants ranging from 22 to 41 years of age. It contains 9 different illumination setups and 3 different filming devices. Comparatively, VIPL has been known for its variation in settings, thus raising difficulty for models to generalize and discard noise. Meanwhile, VIPL presents a typical long-tail data distribution with both sides lacking enough data points. As a result, VIPL is a perfect dataset for us to test our model's ability to both learn physiological patterns through class-imbalanced situation while also generalizing well to a diverse setting range.

In addition to VIPL-HR, we further employ two auxiliary datasets to strengthen our evaluation. UTKFace [61] is a large-scale facial dataset covering individuals aged from 0 to 116 years, annotated with age, gender, and ethnicity. Its inherently long-tailed age distribution makes it a suitable proxy to examine our method's ability to handle class imbalance in visual representations. Yelp Review [52], on the other hand, is a widely used benchmark for sentiment analysis, consisting of millions of user reviews with imbalanced rating labels. While not directly related to physiological signals, it allows us to validate whether our semi-supervised learning framework generalizes well across heterogeneous modalities and label distributions. Together, these datasets complement VIPL-HR by offering diverse yet challenging scenarios for testing robustness and generalization.

4.2. Implementation details

For all the testings in VIPL, we use two labeled regime: 2,500 labels and 20,000 labels out of a net 370,000 labeled spatial-temporal maps (STMaps), generated from face videos, all taken from VIPL dataset at random. The STMap is constructed following the procedure illustrated in Fig. 3, where aligned facial regions are divided into ROI blocks, per-channel average color values are extracted over time, and concatenated into a 2D spatial-temporal representation. For the whole task, we use Unet as the baseline model for training. Since this study focuses only on HR instead of a holistic pack of all physiological signals including BVP, SpO2, etc, the up path of Unet is deprecated and only the down path encoders and the classifier head are preserved. The number of classes k is kept at 133, ranging from 47 to 179 HR. The pseudo-labeling threshold τ is set to 0.9, and the sharpening factor T is set to 0.5. The learning rate for all stages is 0.001. For the beginning encoder warm-up with FixMatch, we run 300 epochs each with 5,000 iterations, each iteration being one batch of labeled and unlabeled data. For the classifier warm-up with TFE, we run 10 epochs of 5000 iterations. Note that only classifier has gradient turned on at this stage. For the combined training, we run another 300 epochs with 5000 iterations. Note that the gradient are still decoupled at this stage.

For UTKFace, we adopt four different labeled sample regimes with 30, 50, 250, and 2000 labeled images, while the remaining samples are treated as unlabeled. This design allows us to examine the performance of our method under extreme scarcity (30, 50 labels), moderate supervision (250 labels), and relatively sufficient supervision (2000 labels). For Yelp Review, we follow a similar setup with two regimes, using 250 and 2000 labeled samples, respectively. This configuration enables us to investigate how the proposed method scales across different magnitudes of available labeled data. For UTKFace, we adopt Wide ResNet-28-2 as the backbone for age estimation, training from scratch to assess its semi-supervised learning capability under imbalanced and limited labels. For Yelp Review, we use BERT-Small initialized with pretrained weights, focusing on evaluating how our framework adapts to textual data in sentiment classification.

4.3. Results on VIPL

Table 1 (left) presents the results on VIPL under two labeled regimes (2,500 and 20,000). Our method consistently surpasses all representative semi-supervised baselines, including Π-Model [21], Mean Teacher [46], CLSS [10], UCVME [11], MixMatch [4], and RankUp [17], demonstrating clear improvements in both accuracy and correla-

Table 1. Results on VIPL and Yelp Review under varying label regimes.

Method	VIPL (Video Remote Physiological Measurement)						Yelp Review (NLP Opinion Mining)					
	Labels = 2500			Labels = 20000			Labels = 250			Labels = 2000		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	R^2↑	SRCC↑	MAE↓	R^2↑	SRCC↑
Supervised	9.717±0.424	13.937±0.635	0.407±0.019	7.751±0.307	11.676±0.559	0.571±0.023	0.723±0.023	0.566±0.019	0.769±0.010	0.581±0.021	0.704±0.016	0.840±0.009
Π-Model	9.425±0.376	13.537±0.608	0.408±0.014	7.559±0.310	11.449±0.563	0.583±0.016	0.730±0.024	0.565±0.019	0.769±0.009	0.580±0.019	0.705±0.013	0.841±0.009
Mean Teacher	9.382±0.392	13.440±0.617	0.444±0.025	7.583±0.294	11.526±0.538	0.577±0.020	0.730±0.024	0.565±0.019	0.769±0.009	0.581±0.021	0.704±0.015	0.840±0.010
CLSS	9.634±0.408	13.577±0.630	0.393±0.023	7.508±0.286	11.371±0.530	0.589±0.016	0.721±0.010	0.543±0.011	0.748±0.002	0.602±0.024	0.639±0.016	0.797±0.011
UCVME	8.819±0.355	12.935±0.614	0.459±0.011	7.249±0.262	10.954±0.517	0.609±0.012	0.775±0.006	0.540±0.005	0.763±0.005	0.593±0.015	0.695±0.009	0.836±0.006
MixMatch	8.567±0.339	12.471±0.583	0.486±0.014	7.141±0.283	10.684±0.499	0.626±0.012	0.886±0.004	0.381±0.008	0.660±0.004	0.774±0.015	0.522±0.008	0.740±0.004
RankUp	8.135±0.314	11.985±0.574	0.518±0.011	6.819±0.270	10.231±0.466	0.634±0.011	0.661±0.018	0.645±0.013	0.829±0.002	0.562±0.020	0.735±0.015	0.859±0.009
Ours	**7.266**±0.301	**9.892**±0.472	**0.563**±0.011	**6.084**±0.233	**8.957**±0.391	**0.685**±0.009	**0.629**±0.010	**0.696**±0.009	**0.849**±0.002	**0.501**±0.014	**0.754**±0.008	**0.873**±0.006
Fully-Supervised	5.248±0.103	8.136±0.214	0.773±0.007	5.248±0.103	8.136±0.214	0.773±0.007	0.418±0.003	0.799±0.002	0.896±0.001	0.418±0.003	0.799±0.002	0.896±0.001

Table 2. Comparison of our method and other methods on UTKFace across varying numbers of labeled samples (30, 50, 250, and 2000). The dataset contains 18,964 training images; the rest of the samples are unlabeled.

Method	UTKFace (Image Age Estimation)											
	Labels = 30			Labels = 50			Labels = 250			Labels = 2000		
	MAE↓	R^2↑	SRCC↑	MAE↓	R^2↑	SRCC↑	MAE↓	R^2↑	SRCC↑	MAE↓	R^2↑	SRCC↑
Supervised	15.02±0.80	0.043±0.025	0.265±0.114	14.13±0.56	0.090±0.092	0.371±0.071	9.42±0.16	0.540±0.014	0.712±0.010	6.28±0.06	0.794±0.004	0.862±0.001
Π-Model	14.26±1.02	0.093±0.050	0.288±0.223	13.82±1.02	0.100±0.086	0.387±0.092	9.45±0.30	0.534±0.030	0.706±0.015	6.31±0.10	0.790±0.006	0.860±0.003
Mean Teacher	14.47±1.23	0.068±0.015	0.307±0.146	13.92±0.20	0.127±0.037	0.423±0.023	8.85±0.25	0.586±0.020	0.745±0.013	6.29±0.03	0.793±0.004	0.862±0.001
CLSS	14.57±0.49	0.047±0.012	0.282±0.113	13.61±0.92	0.138±0.101	0.447±0.074	9.10±0.15	0.586±0.016	0.737±0.014	6.29±0.01	0.794±0.003	0.862±0.001
UCVME	13.76±0.83	0.115±0.078	0.372±0.124	13.49±0.95	0.157±0.110	0.412±0.127	8.63±0.17	0.626±0.006	0.767±0.007	5.90±0.07	0.821±0.007	0.877±0.002
MixMatch	12.50±0.53	0.290±0.026	0.616±0.046	11.44±0.45	0.401±0.028	0.674±0.035	7.95±0.15	0.692±0.013	0.832±0.008	6.03±0.07	0.824±0.004	0.883±0.002
RankUp	11.58±0.55	0.359±0.015	0.606±0.022	9.96±0.62	0.514±0.043	0.703±0.019	7.06±0.11	0.751±0.011	0.835±0.008	5.61±0.07	0.838±0.003	0.887±0.003
Ours	**10.74**±0.43	**0.512**±0.011	**0.728**± 0.033	**7.90**±0.45	**0.598**±0.023	**0.753**±0.017	**5.81**±0.12	**0.815**±0.008	**0.879**±0.007	**5.22**±0.03	**0.861**±0.002	**0.902**±0.001
Fully-Supervised	4.85±0.01	0.875±0.000	0.910±0.001	4.85±0.01	0.875±0.000	0.910±0.001	4.85±0.01	0.875±0.000	0.910±0.001	4.85±0.01	0.875±0.000	0.910±0.001

tion. At 2,500 labels, our method achieves MAE = 7.266, RMSE = 9.892, and r = 0.563, outperforming the strongest baseline RankUp (MAE = 8.135, r = 0.518) by nearly 11% in error reduction and 0.045 in correlation gain. This indicates that our framework more effectively leverages unlabeled data under limited supervision, where model generalization is often most challenging. With 20,000 labels, our method still delivers the best results (MAE = 6.084, RMSE = 8.957, r = 0.685), maintaining a clear advantage even though the performance gaps among baselines typically narrow. Importantly, our approach closes much of the distance to the fully-supervised upper bound (MAE = 5.248, RMSE = 8.136, r = 0.773), suggesting that explicitly addressing imbalance can substantially reduce the need for large-scale labeled data. These consistent improvements stem from three complementary components: (1) the decoupled co-learning strategy, which yields class-agnostic representations less biased toward frequent HR ranges; (2) the Tail-class Feature Enhancement (TFE), which enriches underrepresented HR cases and prevents classifier overfitting to the head; and (3) Label Distribution Smoothing (LDS), which restores the regression-like continuity of HR estimation and stabilizes the training process. Together, these mechanisms enable the model to capture both central and rare HR patterns, yielding robustness across label-scarce and label-abundant settings, and directly addressing the long-tail imbalance that has long hindered rPPG-based heart rate estimation in real-world scenarios.

4.4. Results on UTKFace and Yelp Review

Table 1 (right) and Table 2 further evaluate our method on Yelp Review (text-based sentiment analysis) and UTKFace (image-based age estimation), two datasets that, despite differing in modality, share challenges with rPPG such as long-tailed distributions and limited labeled samples. On UTKFace, our method shows clear improvements in extremely low-label regimes (30 and 50 labels), where competing methods suffer from majority-class dominance and degraded performance, while our framework achieves consistently better MAE and correlation metrics by preserving minority-class signals and mitigating imbalance-induced bias. As the number of labeled samples increases (250 and 2000), our method maintains its lead, confirming that the proposed approach scales effectively across different supervision levels. On Yelp Review, our method surpasses all baselines under both 250 and 2000 labels, demonstrating that the framework generalizes beyond visual physiological tasks to heterogeneous textual data. It is still worth noting that the improvements presented in Yelp Review and UTKFace are not as significant as what is shown in VIPL. This is largely attributed to the lack of VIPL-specific context utilized at data augmentation, losing a particular advantage task-wise. Nevertheless, The consistent improvements across both datasets highlight that integrating CoSSL with

TFE and LDS provides a principled and domain-agnostic solution for CISSL, achieving robust gains across physiological, visual, and textual modalities.

4.5. Comparison with Fully-Supervised Methods on VIPL

As shown in Table 3, we directly compare our semi-supervised framework, which relies only on a limited portion of labeled data, against fully-supervised state-of-the-art (SOTA) methods trained with the entire labeled dataset. Compared with traditional signal decomposition approaches such as SAMC, POS, and CHROM, our method clearly achieves superior results, reducing MAE and RMSE by a large margin and yielding substantially higher correlation scores. When compared with modern deep learning-based methods such as BVPNet, CVD, Physformer, Dual-GAN, NEST, and DOHA, our semi-supervised framework inevitably shows a performance gap due to using fewer annotations, while these fully-supervised approaches exploit the entire dataset. Nevertheless, this gap has been considerably narrowed to within 1.32 in terms of MAE and 0.15 in terms of r, demonstrating that our method approaches state-of-the-art performance despite relying on significantly fewer annotations. This advantage primarily stems from our design that explicitly addresses the long-tail problem in heart rate distribution: by emphasizing tail-class representation and leveraging unlabeled samples through the semi-supervised paradigm, our framework improves robustness in handling elevated and abnormal HR cases that are often underrepresented in training data even for fully-supervised SOTA methods.

Table 3. Comparison of HR estimation results on the VIPL-HR database against fully-supervised state-of-the-art methods using the full labeled dataset. **Bold** denotes the best performance. Symbols: ↑ indicates higher is better, ↓ indicates lower is better.

Method	**MAE↓**	**RMSE↓**	**r↑**
Baseline [32]	5.25	8.14	0.77
SAMC [47]	15.9	21.0	0.11
POS [51]	11.5	17.2	0.30
CHROM [13]	11.4	16.9	0.28
I3D [5]	12.0	15.9	0.07
DeepPhy [7]	11.0	13.8	0.11
BVPNet [12]	5.34	7.85	0.70
CVD [33]	5.02	7.97	0.79
Physformer [59]	4.97	7.79	0.78
Dual-GAN [27]	4.93	7.68	0.81
NEST [28]	4.76	7.51	0.84
DOHA [43]	4.95	7.73	0.80
Ours (Semi-Supervised)	6.08	8.96	0.69

Table 4. Ablation study of TFE and LDS.

Method	MAE↓	R^2↑	SRCC↑
Ours (w/o TFE)	6.452±0.229	9.203±0.403	0.655±0.009
Ours (w/o LDS)	6.410±0.242	9.196±0.413	0.667±0.009
Ours	**6.084**±0.233	**8.957**±0.391	**0.685**±0.009

4.6. Ablation

We further conduct an ablation study to examine the contribution of Tail-class Feature Enhancement (TFE) and Label Distribution Smoothing (LDS) within our framework, as shown in Table 4. Removing either component leads to a clear degradation in performance: without TFE or without LDS, the model shows higher error and weaker correlation compared to the full design. These results highlight the complementary roles of the two modules. TFE directly addresses class imbalance by augmenting samples associated with rare HR ranges, thereby improving the classifier's ability to handle long-tail distributions. LDS, on the other hand, smooths the label distribution to restore the regression-like continuity of HR estimation, mitigating discontinuities introduced by the classification formulation. The combination of TFE and LDS thus provides both better balance across classes and a more faithful modeling of HR continuity, leading to consistent improvements across all metrics. This confirms that tackling both imbalance and continuity is essential for robust semi-supervised rPPG estimation, and validates the effectiveness of our design choices.

5. Conclusion

In this paper, we propose an integration of CoSSL decoupled co-learning with domain-specific explicit priors tackling camera, frame rate, skin, illumination, and motion differences. Through the use of co-learning framework together with TFE, the model are prevented to learn biased feature and classification toward centered range of HR and instead are generalizable to all HR across the big range, including elevated and abnormal cases. The use of LDS further effectively refines the probability of TFE to be used by tackling the continuity problem not represented in CoSSL as a classification task. Overall, the model achieves significant improvement in VIPL dataset, proving the systematic improvement in rPPG and addressing class-imbalanced distribution and generalization problem. Meanwhile, the success of our method in Yelp Review and UTKFace further proves the feasibility of the integration of CoSSL with TFE and LDS in CISSL problem. In future, more ablation studies are needed to closely examine the function and advantage of TFE and LDS in CISSL, and more rigorous empirical experimentations are needed to fully test on the purpose each particular parts of the paradigm served.

References

[1] Bhargav Acharya, William Saakyan, Barbara Hammer, and Hanna Drimalla. Generalization of video-based heart rate estimation methods to low illumination and elevated heart rates. *arXiv preprint arXiv:2503.11697*, 2025. 2

[2] Philip Bachman, Quais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014. 3

[3] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 3, 4

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 3, 4, 5, 6

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. IEEE CVPR*, pages 6299–6308, 2017. 8

[6] Shutao Chen, Sui Kei Ho, Jing Wei Chin, Kin Ho Luo, Tsz Tai Chan, Richard HY So, and Kwan Long Wong. Deep learning-based image enhancement for robust remote photoplethysmography in various illumination scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6085, 2023. 5

[7] 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. 2, 8

[8] Chun-Hong Cheng, Kwan-Long Wong, Jing-Wei Chin, Tsz-Tai Chan, and Richard HY So. Deep learning methods for remote heart rate measurement: A review and future research agenda. *sensors*, 21(18):6296, 2021. 2

[9] Li-Wen Chiu, Yang-Ren Chou, Yi-Chiao Wu, and Bing-Fei Wu. Deep-learning-based remote photoplethysmography measurement in driving scenarios with color and near-infrared images. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023. 1

[10] Weihang Dai, Yao Du, Hanru Bai, Kwang-Ting Cheng, and Xiaomeng Li. Semi-supervised contrastive learning for deep regression with ordinal rankings from spectral seriation. *Advances in Neural Information Processing Systems*, 36:57087–57098, 2023. 3, 6

[11] Weihang Dai, Xiaomeng Li, and Kwang-Ting Cheng. Semi-supervised deep regression with uncertainty consistency and variational model ensembling via bayesian neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7304–7313, 2023. 3, 6

[12] Abhijit Das, Hao Lu, Hu Han, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Bvpnet: Video-to-bvp signal prediction for remote heart rate estimation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021. 3, 8

[13] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE transactions on biomedical engineering*, 60(10):2878–2886, 2013. 2, 8

[14] Uday Debnath and Sungho Kim. A comprehensive review of heart rate measurement using remote photoplethysmography and deep learning. *BioMedical Engineering OnLine*, 24(1): 73, 2025. 1

[15] Yue Fan, Dengxin Dai, Anna Kukleva, and Bernt Schiele. Cossl: Co-learning of representation and classifier for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14574–14584, 2022. 2, 3, 4

[16] Lan-Zhe Guo and Yu-Feng Li. Class-imbalanced semi-supervised learning with adaptive thresholding. In *International conference on machine learning*, pages 8082–8094. PMLR, 2022. 3

[17] Pin-Yen Huang, Szu-Wei Fu, and Yu Tsao. Rankup: Boosting semi-supervised regression with an auxiliary ranking classifier. *Advances in Neural Information Processing Systems*, 37:107444–107468, 2024. 3, 6

[18] Minsung Hyun, Jisoo Jeong, and Nojun Kwak. Class-imbalanced semi-supervised learning. *arXiv preprint arXiv:2002.06815*, 2020. 3

[19] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 2

[20] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. *Advances in neural information processing systems*, 33: 14567–14579, 2020. 3

[21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3, 6

[22] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896. Atlanta, 2013. 3

[23] Hyuck Lee, Seungjae Shin, and Heeyoung Kim. Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. *Advances in Neural Information Processing Systems*, 34:7082–7094, 2021. 3

[24] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jedrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *2011 federated conference on computer science and information systems (FedCSIS)*, pages 405–410. IEEE, 2011. 2

[25] 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. 3

[26] Xin Liu, Yuting Zhang, Zitong Yu, Hao Lu, Huanjing Yue, and Jingyu Yang. rppg-mae: Self-supervised pretraining with masked autoencoders for remote physiological measurements. *IEEE Transactions on Multimedia*, 26:7278–7293, 2024. 3

[27] 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. 3, 8

[28] Hao Lu, Zitong Yu, Xuesong Niu, and Ying-Cong Chen. Neuron structure modeling for generalizable remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18589–18599, 2023. 2, 3, 8

[29] Ewa Magdalena Nowara, Tim K. Marks, Hassan Mansour, and Ashok Veeraraghavan. Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 1

[30] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Asian conference on computer vision*, pages 562–576. Springer, 2018. 6

[31] 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. 1, 3, 5, 6

[32] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Trans. on Image Process.*, 29:2409–2423, 2020. 8

[33] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *Proc. ECCV*, 2020. 8

[34] Sang Bae Park, Gyehyun Kim, Hyun Jae Baek, Jong Hee Han, and Joon Ho Kim. Remote pulse rate measurement from near-infrared videos. *IEEE Signal Processing Letters*, 25(8):1271–1275, 2018. 1

[35] 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. 2

[36] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. *Advances in neural information processing systems*, 28, 2015. 3

[37] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 3

[38] Marko Savic and Guoying Zhao. Physu-net: Long temporal context transformer for rppg with self-supervised pre-training. In *International Conference on Pattern Recognition*, pages 228–243. Springer, 2024. 3

[39] Marko Savic and Guoying Zhao. Rs+ rppg: Robust strongly self-supervised learning for rppg. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2, 3, 5

[40] Hang Shao, Lei Luo, Jianjun Qian, Mengkai Yan, Shuo Chen, and Jian Yang. Remote photoplethysmography in real-world and extreme lighting scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10858–10867, 2025. 2

[41] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 3, 4

[42] Wei Sun, Qing Sun, Hong-Mei Sun, Qi Sun, and Rui-Sheng Jia. Vit-rppg: a vision transformer-based network for remote heart rate estimation. *Journal of Electronic Imaging*, 32(2): 023024–023024, 2023. 3

[43] Weiyu Sun, Xinyu Zhang, Hao Lu, Ying Chen, Yun Ge, Xiaolin Huang, Jie Yuan, and Yingcong Chen. Resolve domain conflicts for generalizable remote physiological measurement. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8214–8224, 2023. 8

[44] 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. 3

[45] Zhaodong Sun and Xiaobai Li. Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5835–5851, 2024. 3

[46] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3, 6

[47] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proc. IEEE CVPR*, pages 2396–2404, 2016. 8

[48] Mark van Gastel, Sander Stuijk, and Gerard de Haan. Motion robust remote-ppg in infrared. *IEEE Transactions on Biomedical Engineering*, 62(5):1425–1433, 2015. 1

[49] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 2

[50] 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. 2

[51] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Trans. Biomed. Eng.*, 64(7):1479–1491, 2017. 8

[52] Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, et al. Usb: A unified semi-supervised learning benchmark for classification. *Advances in Neural Information Processing Systems*, 35:3938–3961, 2022. 6

[53] Yin Wang, Hao Lu, Ying-Cong Chen, Li Kuang, Mengchu Zhou, and Shuiguang Deng. rppg-hiba: Hierarchical balanced framework for remote physiological measurement. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2982–2991, 2024. 2, 3

[54] Bingjie Wu, Zitong Yu, Yiping Xie, Wei Liu, Chaoqi Luo, Yong Liu, and Rick Siow Mong Goh. Semi-rppg: Semi-supervised remote physiological measurement with curriculum pseudo-labeling. *IEEE Transactions on Instrumentation and Measurement*, 2025. 3

[55] Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *Proceedings of the 38th International Conference on Machine Learning*, pages 11842–11851. PMLR, 2021. 2, 3, 4, 5

[56] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 3

[57] 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. 3

[58] Zitong Yu, Xiaobai Li, Pichao Wang, and Guoying Zhao. Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection. *IEEE Signal Processing Letters*, 28:1290–1294, 2021. 3

[59] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4186–4196, 2022. 3, 8

[60] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in neural information processing systems*, 34:18408–18419, 2021. 3, 4

[61] Song Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 6

[62] Yuting Zhang, Hao Lu, Xin Liu, Yingcong Chen, and Kaishun Wu. Advancing generalizable remote physiological measurement through the integration of explicit and implicit prior knowledge. *IEEE Transactions on Image Processing*, 2025. 2, 3, 5

The research project would not have been successfully completed without the guidance and support of my research mentors, so I would like to start by giving the utmost gratitude to Professor Deng, who introduced me to the field of research and this particular project focusing on rPPG, and to Mr. Wang, who has accompanied me throughout my journey in this research project from the very beginning. The entire research process was carried out free of charge.

Starting from the very beginning, Professor Deng introduced me to the field of rPPG and explained how studies have been focusing on pushing the models' capability to a stage where decent generalization abilities could be applied to a wide variety of real-world cases beyond laboratory settings. This would lead to a revolutionary change in how physiological signals are measured, increasing the accessibility of such measurements to many people in need. As such, I started my research in rPPG. It was at this point that Mr. Wang became my teacher in research.

As the sole participant and author of this work, I was primarily responsible for most of the work: starting from reading literature, framing a research idea that was innovative, reading more papers based on my proposed question and topic of interest, reviewing codes open sourced by authors of different studies, adapting and integrating codes from others' work into my own frameworks, writing my own overall training framework, adapting to rPPG datasets, designing experiments, conducting ablation studies, and finally writing the essay.

Professor Deng was mainly responsible for introducing me to the field. Teacher Wang took a more detailed role: he helped me learn how to use the lab server, gave me some baseline model codes and papers unavailable to the public, gave me access to rPPG datasets, helped me correct some conceptual mistakes and code errors, and helped to check for mistakes in my paper draft. Note that because rPPG datasets are sensitive in nature, as they reveal personal information of test subjects, the only way of accessing them is through universities that have the authorization to download from dataset creators. Also, since I was only focusing on rPPG task, Mr. Wang also provided me with some data and testing results of other methods for UTKFace and Yelp Review. Alongside, Professor Deng also provided occasional supports for me when encountered with difficulties or questions in project, ranging from theories to data analysis. He had also been caring for my work experience within the laboratory.

Many problems arose during the process of research. From the very beginning, I had a hard time coming up with innovative ideas to address the generalization issue in rPPG, after reviewing many existing studies in the field. Mr. Wang suggested that I should also look at other studies relevant to rPPG – those that dealt with computer vision tasks in general. Since the task inherently lacks data, I started to look for semi-supervised learning methods for CV studies. It was then that I came across essential concepts of deep regression and class imbalance, which laid the building blocks for my later research. Another example of a problem occurred during the coding phase. Due to my lack of experience in using the server to run Python scripts, I ended up struggling with the hardware usage for my code: I was taking up too many threads and workers during the data loading phase, since this is the phase involving reading and augmenting data samples. Unlike the code I was adapting from, which only used PIL images, rPPG demands more computational and storage resources due to both the large videos and the STMaps. In the end, the problem was resolved after taking a closer examination of the data loading process together with Teacher Wang. I made a separate data processing script and performed as many preprocessing steps as possible before data loading, storing the intermediate outputs in a .npz package to be read into RAM during the actual training, so that the machine would not need to repeatedly read and

process data. Apart from these, there were many more minor issues that were either resolved by me or together with Mr. Wang.

In the end, I want to once again thank Professor Deng and Teacher Wang for the time and effort they dedicated to guiding me in this research. It would not have been possible for me to have such a formal experience of scientific research without them answering my endless questions. In addition, I would also like to thank my friends and parents, whose dedication and kindness helped me balance my research with other aspects of my life.

Below are the resumes for my two teachers:

邓水光简历

浙江大学计算机科学与技术学院教授/博导，求是特聘学者，国家杰出青年基金获得者，IET Fellow / EAI Fellow。现担任浙江大学社会科学研究院副院长、浙江大学中原研究院大数据与人工智能研究中心主任、浙江大学数字农业农村研究中心副主任、浙江省现代服务业电子服务工程技术研究中心副主任。分别于2002、2007年毕业于浙江大学获计算机专业学士和博士学位，于2014年和2015年在美国麻省理工学院、斯坦福大学访问研究。曾获中国青年五四奖章、国家万人计划青年拔尖人才、IEEE TCSVC Rising Star、CCF服务计算杰出成就奖等荣誉称号。主要研究方向为服务计算、边缘计算、软件工程、大数据等。担任IEEE Trans. on Services Computing、Knowledge and Information Systems等国际期刊Associate Editor/Survey Editor。近年来，在国际权威期刊和会议上发表论文100余篇，获9次国际期刊和会议的杰出论文奖/最佳论文奖/最佳学生论文奖，入围2022年全球前2%顶尖科学家终身榜，入选爱思唯尔2020年、2021年中国高被引学者；授权国家发明专利100余项，出版5部学术专著；主持国家杰出青年基金项目、国家重点研发计划项目、国家自然科学基金重点项目等，获国家科技进步二等奖1项，省部科技进步一等奖5项。

教育与工作经历

- 2021.11 ~ 至今，浙江大学社会科学研究院 副院长
- 2016.12 ~ 至今，浙江大学 计算机科学与技术学院 教授
- 2015.01 ~ 2015.12，美国斯坦福大学，访问学者
- 2014.01 ~ 2014.12，美国麻省理工学院，访问学者
- 2009.12 ~ 2016.11，浙江大学 计算机科学与技术学院，副教授
- 2007.07 ~ 2009.11，浙江大学 计算机科学与技术学院，博士后/讲师
- 2002.09 ~ 2007.06，浙江大学 计算机科学与技术学院，博士
- 1998.09 ~ 2002.06，浙江大学 计算机科学与技术学院，本科

主要荣誉与获奖

- 中国电子学会科技进步一等奖 *2022*
- 全球前2%顶尖科学家终身榜 *2022*
- 国家杰出青年基金获得者 *2021*
- CCF服务计算杰出成绩奖 *2020*
- IEEE TII 杰出论文奖 *2020*
- IET Fellow *2020*
- 国家万人计划青年拔尖人才 *2019*
- 浙江省科技进步二等奖 *2018*
- IEEE TCSVC Rising Star Award（年度唯一） *2018*
- ICSOC 2017 唯一最佳论文奖（大陆学者首次） *2017*
- WWW 2017最佳POST提名奖 *2017*
- 浙江省杰出青年基金 *2017*
- 浙江省科技进步一等奖 *2014*
- IEEE SCC 2012最佳学生论文奖（大陆学者首次） *2012*
- 国家科技进步二等奖 *2010*
- 教育部科技进步一等奖 *2008*
- 浙江省科技进步一等奖 *2007*
- 中国“五四”青年奖章 *2007*

王胤

Wang Yin

教育经历

2021.09 至 2026.06 浙江大学 计算机科学与技术 博士
2016.09 至 2019.06 中南大学 软件工程 硕士
2012.09 至 2016.06 周口师范 软件工程 学士

获奖经历

2021.01 至 2022.12 优秀研究生 三好研究生
2023.01 至 2024.12 优秀研究生 CCF2024 中国数字服务大会特邀讲者

工作经历

2019.07 - 2021.08 平安金融壹账通 GammaLab 人工智能研究院，算法工程师
在智能视觉方向具有扎实的研究与工程经验，围绕情绪分析、金融文档识别等核心问题，完成了多个关键模块的算法设计与系统部署。相关成果应用于金融风控、智能面审、文档处理等真实业务场景，涵盖从模型研发到服务端及移动端的系统落地，显著提升了中后台智能系统在复杂环境下的决策效率与识别准确性。

研究方向

半监督学习，长尾学习，表征学习，大模型遗忘学习，边缘智能

学术论文

Yin Wang, Hao Lu, Ying-Cong Chen, Li Kuang, Mengchu Zhou, Shuiguang Deng. rPPG-HiBa: Hierarchical Bal-
anced Framework for Remote Physiological Measurement. Proceedings of the 32nd ACM International Conference on Multimedia. (ACM MM'24) (CCF-A，第一作者)

Yin Wang, Zixuan Wang, Hao Lu, Zhen Qin, Hailiang Zhao, Guanjie Cheng, Xin Du, Ge Su, Li Kuang, Mengchu
Zhou, Shuiguang Deng. SeMi: When Imbalanced Semi-Supervised Learning Meets Mining Hard Examples. Proceed-ings of the 33nd ACM International Conference on Multimedia. (ACM MM'25)(CCF-A，第一作者)

Yin Wang, Zixuan Wang, Hao Lu, Zhen Qin, Hailiang Zhao, Guanjie Cheng, Xin Du, Ge Su, Cheng Zhang, Li Kuang, Mengchu Zhou, Shuiguang Deng. GaussianMatch: Semi-Supervised Regression with Pseudo-Label Filtering via Multi-View Gaussian Consistency. (AAAI'26.) (CCF-A，第一作者, 在投)

Yin Wang, Xiaohang Zhang, Zhen Qin, Guanjie Cheng, Li Kuang, Mengchu Zhou, Shuiguang Deng. From Head to Tail: Revisiting Multi-View Consistency under Imbalanced Semi-Supervised Regression. (WWW'26) (CCF-A，第一作者, 拟投)

Yin Wang, Zhen Qin, Cheng Zhang, Li Kuang, Mengchu Zhou, Shuiguang Deng. EnProbe: Entropy-Guided Probing for Entity-Level Unlearning in Large Language Models. (WWW'26) (CCF-A，第一作者, 拟投)

Zhou Tong, Yin Wang, Xuanwen Bao, Yu Deng, Bo Lin, Ge Su, Kejun Ye, Xiaomeng Dai , Hangyu Zhang, Lulu Liu, Wenyu Wang, Yi Zheng, Weijia Fang, Peng Zhao, Peirong Ding , Shuiguang Deng, Xiangming Xu. Development of a whole-slide-level segmentation-based dMMR/pMMR deep learning detector for colorectal cancer. (Iscience, 2023) (浙江大学第一附属医院合作课题, 学生一作且同等贡献于第一作者,IF=4.1, JCR 一区)

Dalei Jiang, Yin Wang, Feng Zhou, Hongtao Ma, Wenting Zhang, Weijia Fang, Peng Zhao and Zhou Tong. Residual refinement for interactive skin lesion segmentation. (Journal of Biomedical Semantics, 2021) (浙江大学第一附属医院合作课题, 学生一作且同等贡献于第一作者,IF=2.0, JCR 三区)

Yuting Zhang, Hao Lu, Qingyong Hu, Yin Wang, Kaishen Yuan, Xin Liu, Kaishun Wu. Period-LLM: Extending the Periodic Capability of Multimodal Large Language Model. Proceedings of the Computer Vision and Pattern Recognition Conference. (CVPR'25) (CCF-A, 第四作)

Xuanwen Bao, Qiong Li, Dong Chen, Xiaomeng Dai, Chuan Liu,Weihong Tian, Hangyu Zhang, Yuzhi Jin,YinWang, Jinlin Cheng, Chunyu Lai, Chanqi Ye, Shan Xin, Xin Li, Ge Su, Yongfeng Ding, Yangyang Xiong, Jindong Xie, Vincent Tano, Yanfang Wang, Wenguang Fu, Shuiguang Deng, Weijia Fang, Jianpeng Sheng, Jian Ruan, Peng Zhao. A multi-omics analysis-assisted deep learning model identifies a macrophage-oriented module as a potential therapeutic target in colorectal cancer. (Cell Reports Medicine, 2024.) (浙江大学第一附属医院合作项目, 第九作者,IF=10.6, JCR 一区)