Neuron Structure Modeling for Generalizable Remote Physiological Measurement

Hao LU\textsuperscript{1,2}, Zitong YU\textsuperscript{3}, Xuesong NIU\textsuperscript{4}, Yingcong CHEN\textsuperscript{1,2,*}
\textsuperscript{1}The Hong Kong University of Science & Technology (Guangzhou), \textsuperscript{2}The Hong Kong University of Science & Technology, \textsuperscript{3}Great Bay University, \textsuperscript{4}Kuaishou Technology

hlu585@connect.hkust-gz.edu.cn, zitong.yu@ieee.org, nxsedson@gmail.com, yingcongchen@ust.hk

Abstract

Remote photoplethysmography (rPPG) technology has drawn increasing attention in recent years. It can extract Blood Volume Pulse (BVP) from facial videos, making many applications like health monitoring and emotional analysis more accessible. However, as the BVP signal is easily affected by environmental changes, existing methods struggle to generalize well for unseen domains. In this paper, we systematically address the domain shift problem in the rPPG measurement task. We show that most domain generalization methods do not work well in this problem, as domain labels are ambiguous in complicated environmental changes. In light of this, we propose a domain-label-free approach called NEuron STructure modeling (NEST). NEST improves the generalization capacity by maximizing the coverage of feature space during training, which reduces the chance for under-optimized feature activation during inference. Besides, NEST can also enrich and enhance domain invariant features across multi-domain. We create and benchmark a large-scale domain generalization protocol for the rPPG measurement task. Extensive experiments show that our approach outperforms the state-of-the-art methods on both cross-dataset and intra-dataset settings. The codes are available at https://github.com/LuPaoPao/NEST.

1. Introduction

Physiological signals such as heart rate (HR), and heart rate variability (HRV), respiration frequency (RF) are important body indicators that serve not only as vital signs but also track the level of sympathetic activation [17, 33, 54]. Traditional physiological measurements, such as electrocardiograms, heart rate bands, and finger clip devices, have high accuracy. However, they are costly, intrusive, and uncomfortable to wear for a long time. Remote photoplethysmography (rPPG) can extract blood volume pulse (BVP) from face video, which analyzes the periodic changes of the light absorption of the skin caused by heartbeats. Then various physiological indicators (such as HR and HRV) can be calculated based on BVP signals [37, 69, 70]. With non-intrusion and convenience, the rPPG-based physiological measurement method can only use an ordinary camera to monitor physiological indicators and gradually become a research hotspot in the computer vision field [7, 28, 32, 36, 44, 49].

Traditional rPPG measurement methods include signal blind decomposition [21,34,42] and color space transformation [9, 59, 63]. These approaches rely on heartbeat-related...
statistical information, only applicable in constrained environments. In recent years, deep learning (DL) based approaches \cite{6,19,25,28,36,44,50,69,70} have shown their great potentials in rPPG measurement. By learning dedicated rPPG feature representation, these methods achieve promising performance in much more complicated environments \cite{1,36,47,52,68}.

However, deep learning methods suffer from significant performance degradation when applied in real-world scenarios. This is because most training data are captured in lab environments with limited environmental variations. With domain shifts (e.g., different illumination, camera parameters, motions, etc.), these models may struggle to generalize for the unseen testing domain. To validate this, we conduct a cross-dataset evaluation shown in Fig. 1(b). As shown, all DL-based methods do not work well in this evaluation. Furthermore, it is worth noting that DeepPhys \cite{6} and TS-CAN \cite{25} even perform inferior to traditional approach POS \cite{63}.

To improve the performance on the unseen domains, one common practice is to incorporate Domain Generalization (DG) approaches, e.g., encouraging intermediate features to be domain-invariant \cite{12,18,30,38,57,62}. However, as shown in Fig. 1(b) (AD \cite{12} and GroupDRO \cite{38}), the improvements are still quite limited. One reason is that existing DG methods assume that domain labels can be clearly defined (e.g., by data source). Unfortunately, different background scenes, acquisition devices, or even identities could cause very different data distributions in rPPG measurement tasks, and such distribution discrepancy may exist within or cross datasets. In this case, explicitly defining domains is very difficult, and simply treating one dataset as one domain may lead to inferior performance. This is termed as agnostic domain generalization. Besides this, as physiological cues are usually much more subtle than the various noise, the model may overfit in the source domain with the limited training data.

In this paper, we propose the NEural STructure modeling (NEST), a principled solution to the abovementioned problems. The main idea of NEST is to narrow the under-optimized and redundant feature space, align domain invariant features, and enrich discriminative features. Our intuition is as follows. Neural structure refers to the channel activation degree in each convolution layer, which reveals the discriminative feature combination for the specific sample. As the limited variation in a certain restricted domain, there are some spaces that the model seldomly is optimized. Out-of-distribution (OOD) samples may cause abnormal activation in these spaces, which may lead to performance degeneration. Therefore, we regularize the neural structure to encourage the model to be more well-conditioned to avoid abnormal activation caused by OOD samples. Specifically, we propose the NEural STructure Coverage Maximization (NEST-CM) that encourages all neural spaces to be optimized during training, reducing the chance of abnormal activation during testing. Secondly, we propose the NEural STructure Targeted Alignment (NEST-TA) that encourage network suppresses domain variant feature by comparing the samples with the similar physiological information. Thirdly, we propose the NEural STructure Diversity Maximization (NEST-DM) to enrich discriminative features against unseen noise. It should be noted that our approach does not rely on domain labels, which is more applicable in the rPPG measurement task. To summarize, our contributions are listed as follows:

1. We are the first to study the domain shift problem in rPPG measurement, which introduces a new challenge, agnostic domain generalization.
2. We propose the NEural STructure modeling to alleviate domain shift, which is performed by narrowing the under-optimized feature space, and enhancing and enriching domain invariant features.
3. We establish a large-scale domain generalization (DG) benchmark for rPPG measurement, which is the first DG protocol in this task. Extensive experiments in this dataset show the superiority of our approach.

2. Related work

2.1. Remote Physiological Measurement

Remote physiological measurement aims to estimate HR and HRV values by analyzing the chrominance changes in the skin. Traditional methods mainly perform color space transformations or signal decomposition to obtain the BVP signal with high SNR \cite{9,10,42,64}. These methods are designed manually for certain lighting scenes with slight motion and can not generalize to the complex environment in the wild. With powerful modeling capacity, the deep learning (DL) model has been successfully used in remote physiological estimation \cite{6,13,35,36,51,67,69}. Among these methods, various model structures are tailored to extract BVP signals from face videos by aggregating spatial and temporal information \cite{6,37,51,69,70}. Besides, many hand-designed feature maps are used to reduce the computational burden of the network by removing the irrelevant background \cite{13,28,36,45,50}. Some work utilizes the distribution of the target domain to improve the model performance \cite{19,26}, which is not practical without touching the unseen domain in advance. Differently, we study the model generalization on domain generalization protocol for rPPG measurement.

2.2. Domain generalization

Domain generalization (DG), i.e., out-of-distribution generalization, aims to generalize to the unseen domain by training a model on multiple or single source domains \cite{62}. The main solution can be categorized into data manipulation, representation learning, and meta-learning. Data
3. Method

3.1. Overall Framework

In this work, we propose the NEuron STructure modeling (NSET), without using domain label, to alleviate the performance degradation caused by domain shift as illustrated in Fig. 2. Specifically, we first compressed the face video into the spatial-temporal map (STMap) as the input of CNN refer to [36] as shown in Fig. 2 (a). Sequentially, we define an Activated NEuron STructure (NEST) representation that describes the activated degree of different channels of the network as shown in Fig. 2 (b). To alleviate the domain shift, we model a unified NEST distribution by three regularizations: (1) NEuron STructure Coverage Maximization (NEST-CM) is used to narrow under-optimized space that is redundant and unstable as shown in Fig. 2 (c); (2) NEuron STructure Targeted Alignment (NEST-TA) is introduced to align multi-domain feature distribution at instance-level as shown in Fig. 2 (d). (3) Neuron Structure Diversity Maximization (NEST-DM) loss to force the network to explore more features with discriminant ability as shown in Fig. 2 (e).

3.2. NEuron STructure Modeling

NEuron STructure Representation. For the rPPG measurement task, the convolution kernels are adaptively activated to extract physiological features against various noise distribution [27, 56, 65]. However, the facial physiological signals are feeble and can be easily affected by various unseen noises, which make the filters activated abnormally. Therefore, the NEuron STructure (NEST) representation is defined to explore how domain shift affects the model robustness in view of activated feature distribution. Specifically, we firstly average the feature map...
the eigenvalues of the last few bases. This could narrow the under-optimized space as shown in Fig. 2 (c). For a training batch with N samples, NEST representation set of j-th layer \( B_j = \{o_{1,j}; o_{2,j}; \ldots; o_{N,j}\} \in R^{N \times C_j} \) is decomposed to \( S_j \cdot V_j \cdot D_j \) by SVD method as shown in Fig. 4, i.e., \( B_j = S_j \cdot V_j \cdot D_j \), where \( V_j \in R^{N \times C_j} \) and \( N>C_j \). The bases with small eigenvalue values are rare and under-optimized during the training stage. To reduce the occurrence probability of under-optimized NEST caused by domain shift, we stimulate under-optimized bases by maximizing the corresponding eigenvalues:

\[
\mathcal{L}_{CM} = -\mathbb{E}_{j \in U} \sum_{i=1}^{C_j} V'_{j,i} \cdot \lambda_{i,j} < t, \quad (1)
\]

where \( \lambda_{i,j} \) represent the \( i \)-th normalized eigenvalue in \( j \)-th layer’ diagonal matrix \( V_j \), and \( \lambda'_{i,j} \) is the eigenvalue less than the threshold \( t \). Considering different channel numbers in each layer, we design an adaptive threshold \( t \) to select the relatively under-optimized bases, i.e., \( t = \frac{\rho}{C_j} \), where \( \rho \) is a trade-off parameter. Besides, \( C_j \) is the number of eigenvalues of \( j \)-th layer under the threshold, and \( U \) is the number of layers.

**NEuron STructure Targeted Alignment.** Besides the NEST-CM introduced above, we further improve the generalization ability by aligning the NEST representation of different domains. Domain invariant learning has demonstrated its effectiveness in DG problems \([12, 18, 30, 38, 57, 62]\). Some recent research \([11, 16, 29, 60]\) further shows that aligning domain features within the same labels could lead to better performance. However, labels in rPPG are continuous, and it is hard to find samples with identical labels in a training batch.

To cope with this, we propose a novel regularization term called NEuron STructure Alignment (NEST-TA). NEST-TA builds a synthesized prototype \( O'_p \) with the same physiological information as the given targeted sample \( O_t \). Then, the feature within the same labels can be aligned across domains like \([11, 16, 29, 60]\). Specifically, we select the top \( K \) samples closest to the targeted ground truth \( y_t \), i.e., \( \{O'_{t,1}, \ldots, O'_{t,K}\} \). The NEST center of the selected samples \( \overline{O}_t \) is calculated by averaging the NEST presentation of selected samples. Then we calculate the direction vectors between the NEST presentation center of the selected samples \( \overline{O}_t \) and the NEST presentation of each sample \( O'_{t,k} \). Following that, the NEST center of the selected samples \( \overline{O}_t \) is calibrated to the synthesized prototype \( O'_p \).
which is performed by using the Gaussian distribution function to fuse these direction vectors:

\[
O^p_i = \mathcal{G} + \sum_{k=1}^{K} W_{i,k} \cdot (y_{i,k} - y_{i,k}')(O_{i,k} - \mathcal{G}),
\]

\[
W_{i,k} = \frac{G(y_{i,k} - \bar{y}_{i}, \sigma)}{\sum_{k=1}^{K} G(y_{i,k} - \bar{y}_{i}, \sigma)},
\]

where \(G(\cdot, \sigma)\) is the Gaussian distribution with the mean of zero and the standard deviation of \(\sigma\), i.e.,

\[
G(y_{i,k} - \bar{y}_{i}, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_{i,k} - \bar{y}_{i})^2}{2\sigma^2}},
\]

where \(\bar{y}_{i}\) is the mean of the target sample. To align the feature distribution across different domains, NEST-TA uses the cosine similarity \(\text{Cos}(\cdot, \cdot)\) to reduce the variant features between targeted samples \(O_i\) and prototype sample \(O_i'\):

\[
\mathcal{L}_{TA} = \mathbb{E}_{i \in N} \text{Cos}(O_i, O_i').
\]

4.1. Datasets

We elaborately select five datasets (including different motion, camera, and lighting conditions) to curate and benchmark large-scale DG protocol: VIPL-HR [36] have nine scenarios, three RGB cameras, different illumination conditions, and different levels of movement. It is worth mentioning that the BVP signal and video of the dataset do not match in the time dimension, so \(\mathcal{L}_p\) is not used for this dataset. We used the output of HR head as the evaluation result. Besides, we normalize STMap to 30 fps by cubic spline interpolation to solve the problem of the unstable frame rate of video [28, 50, 70].

VAV [47] is designed to collect data with the drastic changes of physiological indicators by simulating ten tasks such as a funny joke, 911 emergency call, and odor experience. It is worth mentioning that there are only heart rate labels in the dataset and no BVP signal labels, so \(\mathcal{L}_p\) is not used for this dataset. We used the output of HR head as the evaluation result.

BUAA [68] is proposed to evaluate the performance of the algorithm against various illumination. We only use data with illumination greater than or equal to 10 lux because underexposed images require special algorithms that are not considered in this article. We used the output of BVP head as the evaluation for BUAA, PURE, and UBFC-rPPG datasets.

PURE [52] contains 60 RGB videos from 10 subjects with six different activities, specifically, sitting still, talking, and four rotating and moving head variations. The BVP signals are down-sampled from 60 to 30 fps with cubic spline interpolation to align with the videos.

UBFC-rPPG [1] containing 42 face videos with sunlight and indoor illumination, the ground-truth BVP signals, and HR values were collected by CMS50E.

4.2. Implementation Details

Training Details. Our proposed method is performed under the Pytorch framework. For the generation of STMap from Video, we use the FAN\(^1\) to detect the 2D landmarks of faces [2], and other steps completely follow [36]. STMap in each dataset is sampled with a time window of 256 with step 10. Following [55, 61], we also design spatial and temporal augmentation for STMap, i.e., shuffle each row of STMap and slide the time window. In addition, color jitter and blur

\[
\mathcal{L}_{overall} = k_1 \mathcal{L}_p + \gamma (k_2 \mathcal{L}_{L1} + k_3 \mathcal{L}_{CM} + k_4 \mathcal{L}_{TA} + k_5 \mathcal{L}_{DM}),
\]

where \(k_1\) - \(k_5\) are five trade-off parameters.

\(^1\)https://github.com/1adrianb/face-alignment
augmentation are used for STMap [5]. Then the STMap $x \in R^{25\times256\times3}$ is resized to $x' \in R^{64\times256\times3}$ as the input of the network. We use the ResNet-18 with removing the max pooling layer as the feature extractor, and the final full connection layer is used as the HR estimation head. At the same time, we designed a BVP signal regression head containing four blocks, each block involves a Conv., transpose, and two Conv. layers with batch normalization and ReLU function. Hyper-parameter $\sigma$, $\rho$ and $\tau$ are set to 5, 0.1, and 0.2 according to the experimental result, and the trade-off parameter $k_1 - k_3$ are set to 1, 0.1, 0.001, 0.1 and 0.01 based on the scale of losses. Adam optimizer with a learning rate of 0.001 is used for training. The batch size and iterations are set to 1024 and 20000.

**Performance Metrics.** Following existing methods [6, 28, 37, 50, 70], standard deviation (SD), mean absolute error (MAE), root mean square error (RMSE), and Pearson’s correlation coefficient ($r$) are used to evaluate the HR estimation. MAE, RMSE, and $r$ are used to evaluate the HRV measurement, including low frequency (LF), high frequency (HF), and LF/HF.

### 4.3. Multi-Source Domain Generalization

#### 4.3.1 HR Estimation

For remote HR estimation, all five datasets (i.e., UBFC [1], PURE [52], BUAA [68], VIPL [36], and V4V [47]) are used to evaluate the model generalization. Specifically, four datasets are used for training, and then the other dataset is used to test. Firstly, we performance the,iPhys ² to evaluate three traditional method (i.e., GREEN [59], CHROM [9], and POS [63]) as shown in Tab. 1. Then, we reproduce five popular DL-based methods (i.e., DeepPhys [6], TS-CAN [25], Rhythmnet [36], and BVPNet [7]) for comparison on the same setting. Subsequently, we directly applied the four DG methods (i.e., AD [12], GroupDRO [38], Coral [53], VREx [18], and NCDG [57]) to Rhythmnet with using DeepDG ³.

### 4.4. The performance of traditional methods.

As shown in Tab. 1, traditional algorithms show relatively acceptable results on some simple scenes, such as BUAA and UBFC, but very poor results on other datasets. It is difficult for these traditional algorithms to extract the physiological information on the V4V dataset, which shows that the traditional methods cannot deal with the noise caused by the complex environment and non-rigid motions.

### The performance of DL-based methods.

For DL-based methods, DeepPhys and TS-CAN even perform worse than traditional methods. Rhythmnet, BVPNet, and Dual-GAN perform much better than both other DL-based methods and traditional methods, which suggests that STMap is the more robust input of CNN than the aligned video. Compared with Rhythmnet, BVPNet and Dual-GAN have not significant improvement on the MSDG protocol. Because BVPNet+

---

² https://github.com/danmcduff/iphys-toolbox
³ https://github.com/jindongwang/transferlearning/tree/master/code/DeepDG

<table>
<thead>
<tr>
<th>Method</th>
<th>UBFC MAE</th>
<th>UBFC RMSE</th>
<th>PURE MAE</th>
<th>PURE RMSE</th>
<th>BUAA MAE</th>
<th>BUAA RMSE</th>
<th>VIPL MAE</th>
<th>VIPL RMSE</th>
<th>V4V MAE</th>
<th>V4V RMSE</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>GREEN [59]</td>
<td>8.02</td>
<td>9.18</td>
<td>10.32</td>
<td>14.27</td>
<td>5.82</td>
<td>7.99</td>
<td>6.09</td>
<td>8.29</td>
<td>5.04</td>
<td>7.12</td>
<td>0.56</td>
</tr>
<tr>
<td>POS [63]</td>
<td>7.35</td>
<td>8.04</td>
<td>9.82</td>
<td>13.44</td>
<td>5.04</td>
<td>7.12</td>
<td>14.59</td>
<td>21.26</td>
<td>17.65</td>
<td>23.22</td>
<td>0.19</td>
</tr>
<tr>
<td>Rhythmnet [36]</td>
<td>5.79</td>
<td>7.91</td>
<td>7.39</td>
<td>10.49</td>
<td>3.58</td>
<td>5.17</td>
<td>8.97</td>
<td>12.16</td>
<td>10.16</td>
<td>14.57</td>
<td>0.34</td>
</tr>
<tr>
<td>Dual-GAN [28]</td>
<td>5.55</td>
<td>7.62</td>
<td>7.24</td>
<td>10.27</td>
<td>3.41</td>
<td>5.23</td>
<td>8.88</td>
<td>11.69</td>
<td>10.04</td>
<td>14.44</td>
<td>0.35</td>
</tr>
<tr>
<td>BVPNet [7]</td>
<td>5.43</td>
<td>7.71</td>
<td>7.23</td>
<td>10.25</td>
<td>3.69</td>
<td>5.48</td>
<td>8.45</td>
<td>11.64</td>
<td>10.01</td>
<td>14.35</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Table 1. HR estimation results on MSDG protocol. It is worth noting that we have removed the GRU mechanism in Rhythmnet because it is an unfair comparison with using GRU to aggregate the temporal information of multiple clips instead of only one clip; * means that these methods use the STMap as the input of CNN and use the STMap augmentation strategy as mentioned in section 4.2; + means that these methods are based on baseline (Rhythmnet without GRU).
Net and Dual-GAN consider the performance of the single domain instead of focusing on solving domain shift. Therefore, Rhythmnet is also selected as a baseline to further analyze the generalization on MSDG protocol.

The performance of DG methods. To further improve the generalization of the model, we apply the existing popular DG methods to baseline (Rhythmnet) as shown in Tab. 1. There is a slight model degradation of AD [12], Coral [53], and GroupDRO [38] on some target datasets. The reason behind this is that these algorithms employ inaccurate domain labeling, specifically, a dataset often with different acquisition devices, lighting, and scene cannot be considered as the same domain. On the contrary, VREx [18] and NCDG [57] can slightly improve the performance of Rhythmnet without utilizing the domain label. Inaccurate domain division will make it difficult for the DG algorithm to effectively align the feature distribution across different domains, which is termed as agnostic domain generalization.

Without the domain labels, NEST experimentally manifested significant improvement on MSDG protocol. For the target dataset BUAA, the results without NEST-CM show the best performance. Because NEST-CM may cause the model degradation with activating meaningless features at the instance level, especially for the BUAA with the slight domain shift.

4.3.2 HRV Estimation

We also conduct experiments for HRV (i.e., LF, HF, and LF/HF) and HR measurement on the PURE, UBFC, and BUAA. The reason for choosing only these three datasets is that VIPL and V4V do not have reliable ground-truth BVP signals, as mentioned in subsection 4.1. We also perform three tradition method (GREEN [59], CHROM [9], POS [63]) by iPhys for comparison with our method as shown in Tab. 2. With a more robust NEST presentation distribution, our method also can get better performance than baseline and other traditional methods in all metrics.

4.4. Single-Source Domain Generalization

For a worst-case scenario, the model is trained on only one dataset to generalize to the unseen domains, namely single-source domain generalization (SSDG). Based on this, we select UBFC-rPPG as the source domain and then select PURE and BUAA as the target domain to further evaluate the generalization of NEST as shown in Tab. 3. As we can see, NEST achieves desired performance by training on only one dataset. The reason is that: (1) NEST-CM can avoid the appearance of abnormal NEST; (2) NEST-TA can find the shared and robust feature of the samples within the same label; (3) NEST-DM can distinguish the boundaries between different samples to improve the generalization.

<table>
<thead>
<tr>
<th>Target</th>
<th>Method</th>
<th>MAE $\downarrow$</th>
<th>RMSE $\downarrow$</th>
<th>r $\uparrow$</th>
<th>MAE $\downarrow$</th>
<th>RMSE $\downarrow$</th>
<th>r $\uparrow$</th>
<th>MAE $\downarrow$</th>
<th>RMSE $\downarrow$</th>
<th>r $\uparrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBFC</td>
<td>GREEN</td>
<td>0.2355 0.2841</td>
<td>0.0924</td>
<td>0.2355 0.2841</td>
<td>0.0924</td>
<td>0.6695 0.9512</td>
<td>0.0467</td>
<td>8.0184 9.1776</td>
<td>0.3634</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CHROM</td>
<td>0.2221 0.2817</td>
<td>0.0698</td>
<td>0.2221 0.2817</td>
<td>0.0698</td>
<td>0.6708 1.0542</td>
<td>0.1054</td>
<td>7.2291 8.9224</td>
<td>0.5123</td>
<td></td>
</tr>
<tr>
<td></td>
<td>POS</td>
<td>0.2364 0.2861</td>
<td>0.1359</td>
<td>0.2364 0.2861</td>
<td>0.1359</td>
<td>0.6515 0.9535</td>
<td>0.1345</td>
<td>7.3539 8.0402</td>
<td>0.8165</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>0.0621 0.0813</td>
<td>0.1873</td>
<td>0.0621 0.0813</td>
<td>0.1873</td>
<td>0.1985 0.2667</td>
<td>0.3043</td>
<td>5.1542 7.4672</td>
<td>0.3634</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NEST</td>
<td>0.0597 0.0782</td>
<td>0.2017</td>
<td>0.0597 0.0782</td>
<td>0.2017</td>
<td>0.2138 0.2824</td>
<td>0.3179</td>
<td>4.7471 6.8876</td>
<td>0.8546</td>
<td></td>
</tr>
</tbody>
</table>

| PURE   | GREEN  | 0.2359 0.3002 | 0.0326 | 0.2359 0.3002 | 0.0326 | 0.6525 0.8932 | 0.0417 | 10.3247 14.2693 | 0.4952          |
|        | CHROM  | 0.2096 0.2751 | 0.1059 | 0.2096 0.2751 | 0.1059 | 0.5404 0.8266 | 0.1173 | 9.7914 12.7568 | 0.3732          |
|        | POS    | 0.1959 0.2571 | 0.1684 | 0.1959 0.2571 | 0.1684 | 0.5373 0.8467 | 0.1433 | 9.8273 13.4414 | 0.3342          |
|        | Baseline | 0.0671 0.0923 | 0.6046 | 0.0671 0.0923 | 0.6046 | 0.2864 0.4184 | 0.5526 | 8.2542 11.1765 | 0.6832          |
|        | NEST   | 0.0635 0.0874 | 0.6422 | 0.0635 0.0874 | 0.6422 | 0.2255 0.3505 | 0.5734 | 7.6889 10.4783 | 0.7255          |

| BUAA   | GREEN  | 0.3472 0.3951 | 0.0871 | 0.3472 0.3951 | 0.0871 | 0.6453 0.8632 | 0.0921 | 5.8231 7.9882 | 0.5624          |
|        | CHROM  | 0.3786 0.3237 | 0.0682 | 0.3786 0.3237 | 0.0682 | 0.6813 0.8836 | 0.0715 | 6.0934 8.2938 | 0.5165          |
|        | POS    | 0.3198 0.3762 | 0.0962 | 0.3198 0.3762 | 0.0962 | 0.6275 0.8424 | 0.1127 | 5.0407 7.1198 | 0.6374          |
|        | Baseline | 0.1451 0.1681 | 0.2891 | 0.1451 0.1681 | 0.2891 | 0.5564 0.6904 | 0.2914 | 3.7852 6.3237 | 0.7492          |
|        | NEST   | 0.1436 0.1665 | 0.2955 | 0.1436 0.1665 | 0.2955 | 0.5514 0.6884 | 0.3004 | 3.3723 5.8806 | 0.7647          |

Table 2. HRV and HR estimation results on the MSDG protocol.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE $\downarrow$</th>
<th>RMSE $\downarrow$</th>
<th>r $\uparrow$</th>
<th>MAE $\downarrow$</th>
<th>RMSE $\downarrow$</th>
<th>r $\uparrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PURE</td>
<td>GREEN</td>
<td>10.32 14.27</td>
<td>0.52</td>
<td>5.82 7.99</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CHROM</td>
<td>9.79 12.76</td>
<td>0.37</td>
<td>6.09 8.29</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>POS</td>
<td>9.82 13.44</td>
<td>0.34</td>
<td>5.04 7.12</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>7.54 10.13</td>
<td>0.73</td>
<td>2.97 3.28</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NEST</td>
<td>6.07 9.06</td>
<td>0.76</td>
<td>2.56 2.73</td>
<td>0.78</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. HR estimation results on SSDG protocol by training on the UBFC-rPPG.
Figure 5. Impacts of the hyperparameter (a) $\rho$, (b) $\sigma$, and (c) $\tau$ of our proposed method.

### 4.5. Intra-Dataset Testing on VIPL

Our proposed method can also improve the effectiveness of the model on intra-dataset testing by reducing the harmful effects of abnormal disturbances. Following [28, 36, 37, 70], we test NEST by using a 5-fold cross-validation protocol on VIPL-HR. Our method compared with ten baseline methods, including the traditional methods (SAMC [58], POS [63] and CHROM [9]) and the DL-based methods (I3D [3], DeepPhy [6], RhythmNet [36], CVD [37], BVPNet [8], Physformer [70], and Dual-GAN [28]). The results of these methods are directly from [28, 70] as shown in Tab. 4.

As shown in Tab. 4, the proposed methods outperform all the methods under all metrics. Compared with the baseline, it also gets a noticeable promotion, i.e., 0.4 in MAE and 0.53 in RMSE. It is worth noting that the VIPL dataset has multiple complex scenes and recording devices and cannot be considered as a single domain, termed the agnostic domain problem. This proves that NEST can effectively alleviate the domain shift without domain labels.

### 4.6. Further Analysis

We conduct further experiments on MSDG protocol (testing on UBFC with training on the other four datasets). **Impact of $\rho$ in NEST-CM.** The hyperparameter $\rho$ determines how many bases need to be strengthened by NEST-CM $\mathcal{L}_{CM}$. As illustrated in Fig. 5 (a), NEST could achieve smaller RMSE when $\rho = 0.1$, indicating that the bases with small eigenvalue require more attention and fine-tuning.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\text{SD}_{\downarrow}$</th>
<th>$\text{MAE}_{\downarrow}$</th>
<th>$\text{RMSE}_{\downarrow}$</th>
<th>$\rho_{\uparrow}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAMC [58]</td>
<td>18.0</td>
<td>15.9</td>
<td>21.0</td>
<td>0.11</td>
</tr>
<tr>
<td>POS [63]</td>
<td>15.3</td>
<td>11.5</td>
<td>17.2</td>
<td>0.30</td>
</tr>
<tr>
<td>CHROM [9]</td>
<td>15.1</td>
<td>11.4</td>
<td>16.9</td>
<td>0.28</td>
</tr>
<tr>
<td>I3D [3]</td>
<td>15.9</td>
<td>12.0</td>
<td>15.9</td>
<td>0.07</td>
</tr>
<tr>
<td>DeepPhy [6]</td>
<td>13.6</td>
<td>11.0</td>
<td>13.8</td>
<td>0.11</td>
</tr>
<tr>
<td>BVPNet [8]</td>
<td>7.8</td>
<td>5.34</td>
<td>7.85</td>
<td>0.70</td>
</tr>
<tr>
<td>RhythmNet [36]</td>
<td>8.11</td>
<td>5.30</td>
<td>8.14</td>
<td>0.76</td>
</tr>
<tr>
<td>CVD [37]</td>
<td>7.92</td>
<td>5.02</td>
<td>7.97</td>
<td>0.79</td>
</tr>
<tr>
<td>Physformer [70]</td>
<td>7.74</td>
<td>4.97</td>
<td>7.79</td>
<td>0.78</td>
</tr>
<tr>
<td>Dual-GAN [28]</td>
<td>7.63</td>
<td>4.93</td>
<td>7.68</td>
<td>0.81</td>
</tr>
<tr>
<td>Baseline</td>
<td>8.04</td>
<td>5.21</td>
<td>8.07</td>
<td>0.77</td>
</tr>
<tr>
<td>NEST</td>
<td>7.49</td>
<td>4.76</td>
<td>7.51</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table 4. HR estimation results by our method and several state-of-the-art methods on the VIPL-HR database.

Figure 6. The pearson’s correlation coefficient of the NEST representation of the samples with different labels for the anchor HR 60.

**Impact of $\sigma$ in NEST-TA.** The hyperparameter $\sigma$ tradeoffs the contribution of each selected sample to form the synthesized prototype. As illustrated in Fig. 5, NEST could achieve the best performance with $\sigma = 5$, which shows that the appropriate fusion of the nearby sample is conducive to improving model generalization. The specific reason behind this is that: (1) the synthesized prototype based on multi-domain samples can be used to enhance the domain invariant feature by NEST-TA; (2) NEST-TA can effectively aggregate the label relationship into NEST presentation distribution, which is performed by improving the similarity of NEST representation between the adjacent label samples.

To further clarify the second reason, we calculate the Pearson’s correlation coefficient of the NEST representation of the samples with different labels for the anchor HR 60, as shown in Fig. 6. It is obvious that the distribution changes with NEST-TA loss are more smooth across the continuous HR labels, which allows OOD samples with slight domain shifts to predict nearby heart rates around the ground truth HR value instead of other abnormal HR values.

**Impact of $\tau$ in NEST-DM.** The hyperparameter $\tau$ determines the pushing strength of the different samples. With the smallest $\tau$, NEST-DM $\mathcal{L}_{DM}$ has the overemphasized constraints to distinguish the NEST representation between different samples, which causes a harmful impact on the discriminant ability of the model. Overall, NEST-DM $\mathcal{L}_{DM}$ is an effective way to improve the generalization of model.

### 5. Conclusion

This paper presents the NEural STructure (NEST) modeling, a novel framework that addresses the domain generalization problem in rPPG measurement for the first time. Our framework contains three regularization terms defined in the view of neural structure, which improves the generalization capacity in the different aspects. All three terms do not rely on domain labels, which are especially suitable in rPPG, as defining domains is difficult in this task. To compare the generalization of the existing algorithm and our method, we establish the first DG protocol for this task. We experimentally manifested the substantial improvement of our method on both MSDG and SSDG protocols.
References


