

PhysFormer: Facial Video-based Physiological Measurement with Temporal Difference Transformer

Zitong Yu¹, Yuming Shen², Jingang Shi³, Hengshuang Zhao^{4,2}, Philip Torr², Guoying Zhao^{1*}

¹CMVS, University of Oulu

³Xi'an Jiaotong University

²TVG, University of Oxford

⁴The University of Hong Kong

Abstract

Remote photoplethysmography (rPPG), which aims at measuring heart activities and physiological signals from facial video without any contact, has great potential in many applications. Recent deep learning approaches focus on mining subtle rPPG clues using convolutional neural networks with limited spatio-temporal receptive fields, which neglect the long-range spatio-temporal perception and interaction for rPPG modeling. In this paper, we propose the PhysFormer, an end-to-end video transformer based architecture, to adaptively aggregate both local and global spatio-temporal features for rPPG representation enhancement. As key modules in PhysFormer, the temporal difference transformers first enhance the quasi-periodic rPPG features with temporal difference guided global attention, and then refine the local spatio-temporal representation against interference. Furthermore, we also propose the label distribution learning and a curriculum learning inspired dynamic constraint in frequency domain, which provide elaborate supervisions for PhysFormer and alleviate overfitting. Comprehensive experiments are performed on four benchmark datasets to show our superior performance on both intra- and cross-dataset testings. One highlight is that, unlike most transformer networks needed pretraining from large-scale datasets, the proposed PhysFormer can be easily trained from scratch on rPPG datasets, which makes it promising as a novel transformer baseline for the rPPG community. The codes are available at <https://github.com/ZitongYu/PhysFormer>.

1. Introduction

Physiological signals such as heart rate (HR), respiration frequency (RF), and heart rate variability (HRV) are important vital signs to be measured in many circumstances, especially for healthcare or medical purposes. Traditionally, the Electrocardiography (ECG) and Photoplethysmograph (PPG) are the two most common ways for measuring heart

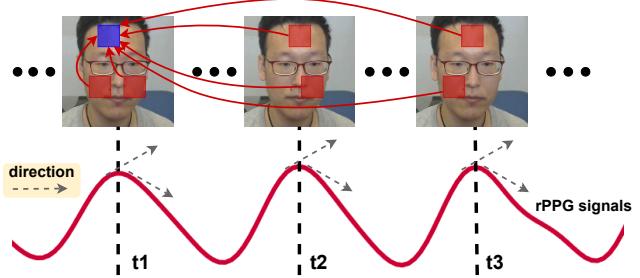


Figure 1. The trajectories of rPPG signals around t_1 , t_2 , and t_3 share similar properties (e.g., trends with rising edge first then falling edge later, and relatively high magnitudes) induced by skin color changes. It inspires the long-range spatio-temporal attention (e.g., blue tube around t_1 interacted with red tubes from intra- and inter-frames) according to their local temporal difference features for quasi-periodic rPPG enhancement. Here ‘tube’ indicates the same regions across short-time consecutive frames.

activities and corresponding physiological signals. However, both ECG and PPG sensors need to be attached to body parts, which may cause discomfort and are inconvenient for long-term monitoring. To counter for this issue, remote photoplethysmography (rPPG) [12, 36, 66] methods are developing fast in recent years, which aim to measure heart activity remotely without any contact.

In earlier studies of facial rPPG measurement, most methods analyze subtle color changes on facial regions of interest (ROI) with classical signal processing approaches [30, 49, 50, 55, 57]. Besides, there are a few color subspace transformation methods [13, 59] which utilize all skin pixels for rPPG measurement. Based on the prior knowledge from traditional methods, a few learning based approaches [25, 44, 45, 51] are designed as non-end-to-end fashions. ROI based preprocessed signal representations (e.g., time-frequency map [25] and spatio-temporal map [44, 45]) are generated first, and then learnable models could capture rPPG features from these maps. However, these methods need the strict preprocessing procedure and neglect the global contextual clues outside the pre-defined ROIs. Meanwhile, more and more end-to-end deep learning based rPPG methods [11, 34, 53, 65, 67] are developed,

*Corresponding author

which treat facial video frames as input and predict rPPG and other physiological signals directly. However, pure end-to-end methods are easily influenced by the complex scenarios (e.g., with head movement and various illumination conditions) and rPPG-unrelated features can not be ruled out in learning, resulting in large performance decrease [63] in realistic datasets (e.g., VIPL-HR [45]).

Recently, due to its excellent long-range attentional modeling capacities in solving sequence-to-sequence issues, transformer [22, 32] has been successfully applied in many artificial intelligence tasks such as natural language processing (NLP) [56], image [15] and video [3] analysis. Similarly, rPPG measurement from facial videos can be treated as a video sequence to signal sequence problem, where the long-range contextual clues should be exploited for semantic modeling. As shown in Fig. 1, rPPG clues from different skin regions and temporal locations (e.g., signal trajectories around t_1 , t_2 , and t_3) share similar properties (e.g., trends with rising edge first then falling edge later and relative high magnitudes), which can be utilized for long-range feature modeling and enhancement. However, different from the most video tasks aiming at huge motion representation, facial rPPG measurement focuses on capturing subtle skin color changes, which makes it challenging for global spatio-temporal perception. Furthermore, video-based rPPG measurement is usually a long-time monitoring task, and it is challenging to design and train transformers with long video sequence inputs.

Motivated by the discussions above, we propose an end-to-end video transformer architecture, namely PhysFormer, for remote physiological measurement. On one hand, the cascaded temporal difference transformer blocks in PhysFormer benefit the rPPG feature enhancement via global spatio-temporal attention based on the fine-grained temporal skin color differences. On the other hand, to alleviate the interference-induced overfitting issue and complement the weak temporal supervision signals, elaborate supervision in frequency domain is designed, which helps PhysFormer learn more intrinsic rPPG-aware features.

The contributions of this work are as follows:

- We propose the PhysFormer, which mainly consists of a powerful video temporal difference transformer backbone. To our best knowledge, it is the first time to explore the long-range spatio-temporal relationship for reliable rPPG measurement.
- We propose an elaborate recipe to supervise PhysFormer with label distribution learning and curriculum learning guided dynamic loss in frequency domain to learn efficiently and alleviate overfitting.
- We conduct intra- and cross-dataset testings and show that the proposed PhysFormer achieves superior or on par state-of-the-art performance without pretraining on large-scale datasets like ImageNet-21K.

2. Related Work

Remote physiological measurement. An early study of rPPG-based physiological measurement was reported in [57]. Plenty of traditional hand-crafted approaches have been developed on this field since then. Selective merging information from different color channels [30, 49, 50] or different ROIs [27, 30] are proven to be efficient for subtle rPPG signal recovery. To improve the signal-to-noise-ratio of the recovered rPPG signals, several signal decomposition methods such as independent component analysis (ICA) [27, 49, 50] and matrix completion [55] are also proposed. In recent years, deep learning based approaches dominate the field of rPPG measurement due to the strong spatio-temporal representation capabilities. On one hand, facial ROI based spatial-temporal signal maps [40, 41, 44, 46, 47] are developed, which alleviate the interference from non-skin regions. Based on these signal maps, 2D-CNNs are utilized for rPPG feature extraction. On the other hand, end-to-end spatial networks [11, 53] and spatio-temporal models [20, 34, 35, 48, 63, 65, 67] are developed, which could recover rPPG signals from the facial video directly. However, previous methods only consider the spatio-temporal rPPG features from adjacent frames and neglect the long-range relationship among quasi-periodic rPPG features.

Transformer for vision tasks. Transformer [32] is proposed in [56] to model sequential data in the field of NLP. Then vision transformer (ViT) [15] is proposed recently by feeding transformer with sequences of image patches for image classification. Many other ViT variants [8, 14, 22, 23, 26, 38, 54, 60, 70] are proposed from then, which achieve promising performance compared with its counterpart CNNs for image analysis tasks [6, 24, 74]. Recently, some works introduce vision transformer for video understanding tasks such as action recognition [1, 3, 4, 16, 21, 39, 42], action detection [37, 58, 62, 73], video super-resolution [5], video inpainting [33, 71], and 3D animation [9, 10]. Some works [21, 42] conduct temporal contextual modeling with transformer based on single-frame features from pretrained 2D networks, while other works [1, 3, 4, 16, 39] mine the spatio-temporal attentions via video transformer directly. Most of these works are incompatible for long-video-sequence (>150 frames) signal regression task. There are two related works [35, 64] using ViT for rPPG feature representation. TransRPPG [64] extracts rPPG features from the preprocessed signal maps via ViT for face 3D mask presentation attack detection [68]. Based on the temporal shift networks [31, 34], EfficientPhys-T [35] adds several swin transformer [38] layers for global spatial attention. Different from these two works, the proposed PhysFormer is an end-to-end video transformer, which is able to capture long-range spatio-temporal attentional rPPG features from facial video directly.

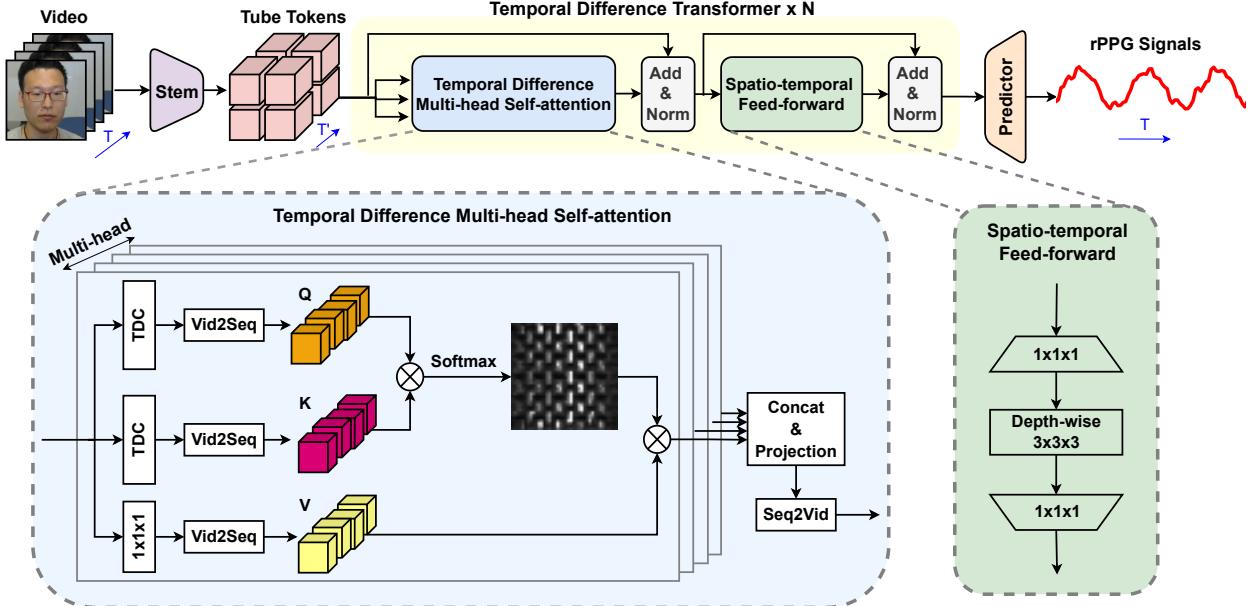


Figure 2. Framework of the PhysFormer. It consists of a shallow stem, a tube tokenizer, several temporal difference transformers, and a rPPG predictor head. The temporal difference transformer is formed from the Temporal Difference Multi-head Self-attention (TD-MHSA) and Spatio-temporal Feed-forward (ST-FF) modules, which enhances the global and local spatio-temporal representation, respectively. ‘TDC’ is short for the temporal difference convolution [63, 69].

3. Methodology

We will first introduce the architecture of PhysFormer in Sec. 3.1, then introduce label distribution learning for rPPG measurement in Sec. 3.2, and at last present the curriculum learning guided dynamic supervision in Sec. 3.3.

3.1. PhysFormer

As illustrated in Fig. 2, PhysFormer consists of a shallow stem \mathbf{E}_{stem} , a tube tokenizer \mathbf{E}_{tube} , N temporal difference transformer blocks $\mathbf{E}_{\text{trans}}^i$ ($i = 1, \dots, N$) and a rPPG predictor head. Inspired by the study in [61], we adopt a shallow stem to extract coarse local spatio-temporal features, which benefits the fast convergence and clearer subsequent global self-attention. Specifically, the stem is formed by three convolutional blocks with kernel size (1x5x5), (3x3x3) and (3x3x3), respectively. Each convolution operator is cascaded with a batch normalization (BN), ReLU and MaxPool. The pooling layer only halves the spatial dimension. Therefore, given an RGB facial video input $X \in \mathbb{R}^{3 \times T \times H \times W}$, the stem output $X_{\text{stem}} = \mathbf{E}_{\text{stem}}(X)$, where $X_{\text{stem}} \in \mathbb{R}^{D \times T \times H/8 \times W/8}$, and D, T, W, H indicate channel, sequence length, width, height, respectively. Then X_{stem} would be partitioned into spatio-temporal tube tokens $X_{\text{tube}} \in \mathbb{R}^{D \times T' \times H' \times W'}$ via the tube tokenizer \mathbf{E}_{tube} . Subsequently, the tube tokens will be forwarded with N temporal difference transformer blocks and obtain the global-local refined rPPG features X_{trans} , which has the same dimensions with X_{tube} . Finally, the rPPG predictor head temporally up-samples, spatially averages, and projects the features X_{trans}

to 1D signal $Y \in \mathbb{R}^T$.

Tube tokenization. Here the coarse feature X_{stem} would be partitioned into non-overlapping tube tokens via $\mathbf{E}_{\text{tube}}(X_{\text{stem}})$, which aggregates the spatio-temporal neighbor semantics and reduces computational costs for the subsequent transformers. Specifically, with the targeted tube size $T_s \times H_s \times W_s$ (the same as the partition step size in non-overlapping setting), the tube token map $X_{\text{tube}} \in \mathbb{R}^{D \times T' \times H' \times W'}$ has length, height and width

$$T' = \left\lfloor \frac{T}{T_s} \right\rfloor, H' = \left\lfloor \frac{H/8}{H_s} \right\rfloor, W' = \left\lfloor \frac{W/8}{W_s} \right\rfloor. \quad (1)$$

Please note that there are no position embeddings after the tube tokenization as the stem at early stage already captures relative spatio-temporal positions.

Temporal difference multi-head self-attention. In self-attention mechanism [15, 56], the relationship between the tokens is modeled by the similarity between the projected query-key pairs, yielding the attention score. Instead of point-wise linear projection, we utilize temporal difference convolution (TDC) [63, 69] for query (Q) and key (K) projection, which could capture fine-grained local temporal difference features for subtle color change description. TDC with learnable w can be formulated as

$$\text{TDC}(x) = \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n) + \theta \cdot (-x(p_0) \cdot \sum_{p_n \in \mathcal{R}'} w(p_n))}_{\text{vanilla 3D convolution}} \quad (2)$$

where p_0 , \mathcal{R} and \mathcal{R}' indicate the current spatio-temporal location, sampled local (3x3x3) neighborhood and sampled adjacent neighborhood, respectively. Then query and key are projected as

$$Q = \text{BN}(\text{TDC}(X_{\text{tube}})), K = \text{BN}(\text{TDC}(X_{\text{tube}})). \quad (3)$$

For the value (V) projection, point-wise linear projection without BN is utilized. Then $Q, K, V \in \mathbb{R}^{D \times T' \times H' \times W'}$ are flattened into sequence, and separated into h heads ($D_h = D/h$ for each head). For the i -th head ($i \leq h$), the self-attention (SA) can be formulated

$$\text{SA}_i = \text{Softmax}(Q_i K_i^T / \tau) V_i, \quad (4)$$

where τ controls the sparsity. We find that the default setting $\tau = \sqrt{D_h}$ in [15, 56] performs poorly for rPPG measurement. According to the periodicity of rPPG features, we use smaller τ value to obtain sparser attention activation. The corresponding study can be found in Table 6. The output of TD-MHSA is the concatenation of SA from all heads and then with a linear projection $U \in \mathbb{R}^{D \times D}$

$$\text{TD-MHSA} = \text{Concat}(\text{SA}_1; \text{SA}_2; \dots; \text{SA}_h) U. \quad (5)$$

As illustrated in Fig. 2, residual connection and layer normalization (LN) would be conducted after TD-MHSA.

Spatio-temporal feed-forward. The vanilla feed-forward network consists of two linear transformation layers, where the hidden dimension D' between two layers is expanded to learn a richer feature representation. In contrast, we introduce a depthwise 3D convolution (with BN and nonlinear activation) between these two layers with extra slight computational cost but remarkable performance improvement. The benefits are two-fold: 1) as a complementation of TD-MHSA, ST-FF could refine the local inconsistency and parts of noisy features; 2) richer locality provides TD-MHSA sufficient relative position cues.

3.2. Label Distribution Learning

Similar to the facial age estimation task [18, 19] that faces at close ages look quite similar, facial rPPG signals with close HR values usually have similar periodicity. Inspired by this observation, instead of considering each facial video as an instance with one label (HR), we regard each facial video as an instance associated with a label distribution. The label distribution covers a certain number of class labels, representing the degree that each label describes the instance. Through this way, one facial video can contribute to both targeted HR value and its adjacent HRs.

To consider the similarity information among HR classes during the training stage, we model the rPPG-based HR estimation problem as a specific L -class multi-label classification problem, where $L=139$ in our case (each integer HR value within [42, 180] bpm as a class). A label distribution $\mathbf{p} = \{p_1, p_2, \dots, p_L\} \in \mathbb{R}^L$ is assigned to each facial

video X . It is assumed that each entry of \mathbf{p} is a real value in the range $[0, 1]$ such that $\sum_{k=1}^L p_k = 1$. We consider the Gaussian distribution function, centred at the ground truth HR label Y_{HR} with the standard deviation σ , to construct the corresponding label distribution \mathbf{p} .

$$p_k = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(k - (Y_{\text{HR}} - 41))^2}{2\sigma^2}\right). \quad (6)$$

The label distribution loss can be formulated as $\mathcal{L}_{\text{LD}} = \text{KL}(\mathbf{p}, \text{Softmax}(\hat{\mathbf{p}}))$, where divergence measure $\text{KL}(\cdot)$ denotes the Kullback-Leibler (KL) divergence [17], and $\hat{\mathbf{p}}$ is the power spectral density (PSD) of predicted rPPG signals.

Please note that the previous work [43] also considers the distribution learning for HR estimation. However, it is totally different with our work: 1) the motivation in [43] is to smooth the temporal HR outliers caused by facial movements across continuous video clips, while our work is more generic, aiming at efficient feature learning across adjacent labels under limited-scale training data; 2) the technique used in [43] is after a post-HR-estimation for the hand-crafted rPPG signals, while our work is to design a reasonable supervision signal \mathcal{L}_{LD} for PhysFormer.

3.3. Curriculum Learning Guided Dynamic Loss

Curriculum learning [2], as a major machine learning regime with philosophy of easy-to-hard curriculum, is utilized to train PhysFormer. In the rPPG measurement task, the supervision signals from temporal domain (e.g., mean square error loss [11], negative Pearson loss [65, 67]) and frequency domain (e.g., cross-entropy loss [46, 63], signal-to-noise ratio loss [53]) provide different extents of constraints for model learning. The former one gives signal-trend-level constraints, which is straightforward and easy for model convergence but overfitting after that. In contrast, the latter one with strong constraints on frequency domain enforces the model learning periodic features within target frequency bands, which is hard to converge well due to the realistic rPPG-irrelevant noise. Inspired by the curriculum learning, we propose the dynamic supervision to gradually enlarge the frequency constraints, which alleviates the overfitting issue and benefits the intrinsic rPPG-aware feature learning gradually. Specifically, exponential increment strategy is adopted, and comparison with other dynamic strategies (e.g., linear increment) will be shown in Table 7. The dynamic loss $\mathcal{L}_{\text{overall}}$ can be formulated as

$$\begin{aligned} \mathcal{L}_{\text{overall}} &= \underbrace{\alpha \cdot \mathcal{L}_{\text{time}}}_{\text{temporal}} + \underbrace{\beta \cdot (\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{LD}})}_{\text{frequency}}, \\ \beta &= \beta_0 \cdot (\eta^{(\text{Epoch}_{\text{current}} - 1)/\text{Epoch}_{\text{total}}}), \end{aligned} \quad (7)$$

where hyperparameters α , β_0 and η equal to 0.1, 1.0 and 5.0, respectively. Negative Pearson loss [65, 67] and frequency cross-entropy loss [46, 63] are adopted as $\mathcal{L}_{\text{time}}$ and

\mathcal{L}_{CE} , respectively. With the dynamic supervision, PhysFormer could perceive better signal trend at the beginning while such perfect warming up facilitates the gradually stronger frequency knowledge learning later.

4. Experimental Evaluation

Experiments of rPPG-based physiological measurement for three types of physiological signals, i.e., heart rate (HR), heart rate variability (HRV), and respiration frequency (RF), are conducted on four benchmark datasets (VIPL-HR [45], MAHNOB-HCI [52], MMSE-HR [55], and OBF [29]).

4.1. Datasets and Performance Metrics

VIPL-HR [45] is a large-scale dataset for remote physiological measurement under less-constrained scenarios. It contains 2,378 RGB videos of 107 subjects recorded with different head movements, lighting conditions and acquisition devices. **MAHNOB-HCI** [52] is one of the most widely used benchmark for remote HR measurement evaluations. It includes 527 facial videos of with 61 fps framerate and 780x580 resolution from 27 subjects. **MMSE-HR** [55] is a dataset including 102 RGB videos from 40 subjects, and the raw resolution of each video is at 1040x1392. **OBF** [29] is a high-quality dataset for remote physiological signal measurement. It contains 200 five-minute-long RGB videos with 60 fps framerate recorded from 100 healthy adults.

Average HR estimation task is evaluated on all four datasets while HRV and RF estimation tasks on high-quality OBF [29] dataset. Specifically, we follow existing methods [41, 46, 67] and report low frequency (LF), high frequency (HF), and LF/HF ratio for HRV and RF estimation. We report the most commonly used performance metrics for evaluation, including the standard deviation (SD), mean absolute error (MAE), root mean square error (RMSE), and Pearson’s correlation coefficient (r).

4.2. Implementation Details

Our proposed method is implemented with Pytorch. For each video clip, we use the MTCNN face detector [72] to crop the enlarged face area at the first frame and fix the region through the following frames. The videos in MAHNOB-HCI and OBF are downsampled to 30 fps for efficiency. The settings $N=12$, $h=4$, $D=96$, $D'=144$ are used for PhysFormer while $\theta=0.7$ and $\tau=2.0$ for TD-MHSA. The targeted tube size $T_s \times H_s \times W_s$ equals to $4 \times 4 \times 4$. In the training stage, we randomly sample RGB face clips with size $160 \times 128 \times 128$ ($T \times H \times W$) as model inputs. Random horizontal flipping and temporally up/down-sampling [63] are used for data augmentation. The PhysFormer is trained with Adam optimizer and the initial learning rate and weight decay are $1e-4$ and $5e-5$, respectively. We cannot find obvious performance improvement using AdamW optimizer. We train models with 25 epochs with fixed setting $\alpha=0.1$

Table 1. Intra-dataset testing results on VIPL-HR [45]. The symbols \blacktriangle , \blacklozenge and \star denote traditional, non-end-to-end learning based and end-to-end learning based methods, respectively. Best results are marked in **bold** and second best in underline.

Method	SD ↓ (bpm)	MAE ↓ (bpm)	RMSE ↓ (bpm)	$r \uparrow$
Tulyakov2016 [55] \blacktriangle	18.0	15.9	21.0	0.11
POS [59] \blacktriangle	15.3	11.5	17.2	0.30
CHROM [13] \blacktriangle	15.1	11.4	16.9	0.28
RhythmNet [45] \blacklozenge	8.11	5.30	8.14	0.76
ST-Attention [47] \blacklozenge	7.99	5.40	7.99	0.66
NAS-HR [40] \blacklozenge	8.10	5.12	8.01	<u>0.79</u>
CVD [46] \blacklozenge	7.92	5.02	7.97	<u>0.79</u>
Dual-GAN [41] \blacklozenge	<u>7.63</u>	4.93	7.68	0.81
I3D [7] \star	15.9	12.0	15.9	0.07
PhysNet [65] \star	14.9	10.8	14.8	0.20
DeepPhys [11] \star	13.6	11.0	13.8	0.11
AutoHR [63] \star	8.48	5.68	8.68	0.72
PhysFormer (Ours) \star	<u>7.74</u>	<u>4.97</u>	<u>7.79</u>	0.78

for temporal loss while exponentially increased parameter $\beta \in [1, 5]$ for frequency losses. We set $\sigma=1.0$ for label distribution learning. The batch size is 4 on one 32G V100 GPU. In the testing stage, similar to [45], we uniformly separate 30-second videos into three short clips with 10 seconds, and then video-level HR is calculated via averaging HRs from three short clips.

4.3. Intra-dataset Testing

HR estimation on VIPL-HR. In these experiments, we follow [45] and use a subject-exclusive 5-fold cross-validation protocol on VIPL-HR. As shown in Table 1, all three traditional methods (Tulyakov2016 [55], POS [59] and CHROM [13]) perform poorly due to the complex scenarios (e.g., large head movement and various illumination) in the VIPL-HR dataset. Similarly, the existing end-to-end learning based methods (e.g., PhysNet [65], DeepPhys [11], and AutoHR [63]) predict unreliable HR values with large RMSE compared with non-end-to-end learning approaches (e.g., RhythmNet [45], ST-Attention [47], NAS-HR [40], CVD [46], and Dual-GAN [41]). Such the large performance margin might be caused by the coarse and overfitted rPPG features extracted from the end-to-end models. In contrast, all five non-end-to-end methods first extract fine-grained signal maps from multiple facial ROIs, and then more dedicated rPPG clues would be extracted via the cascaded models. Without strict and heavy preprocessing procedure in [40, 41, 45–47], our proposed PhysFormer can be trained from scratch on facial videos directly, and achieves comparable performance with state-of-the-art non-end-to-end learning based method Dual-GAN [41]. It indicates that PhysFormer is able to learn the intrinsic and periodic rPPG-aware features automatically.

Table 2. Performance comparison of HR and RF measurement as well as HRV analysis on OBF [29].

Method	HR(bpm)			RF(Hz)			LF(u.n)			HF(u.n)			LF/HF		
	SD	RMSE	<i>r</i>	SD	RMSE	<i>r</i>									
ROI_green [29]▲	2.159	2.162	0.99	0.078	0.084	0.321	0.22	0.24	0.573	0.22	0.24	0.573	0.819	0.832	0.571
CHROM [13]▲	2.73	2.733	0.98	0.081	0.081	0.224	0.199	0.206	0.524	0.199	0.206	0.524	0.83	0.863	0.459
POS [59]▲	1.899	1.906	0.991	0.07	0.07	0.44	0.155	0.158	0.727	0.155	0.158	0.727	0.663	0.679	0.687
CVD [46]◆	1.257	1.26	0.996	0.058	0.058	0.606	0.09	0.09	0.914	0.09	0.09	0.914	0.453	0.453	0.877
rPPGNet [67]★	1.756	1.8	0.992	0.064	0.064	0.53	0.133	0.135	0.804	0.133	0.135	0.804	0.58	0.589	0.773
PhysFormer (Ours)★	0.804	0.804	0.998	0.054	0.054	0.661	0.085	0.086	0.912	0.085	0.086	0.912	0.389	0.39	0.896

Table 3. Intra-dataset results on MAHNOB-HCI [52].

Method	SD ↓ (bpm)	MAE ↓ (bpm)	RMSE ↓ (bpm)	<i>r</i> ↑
Poh2011 [49]▲	13.5	-	13.6	0.36
CHROM [13]▲	-	13.49	22.36	0.21
Li2014 [30]▲	6.88	-	7.62	0.81
Tulyakov2016 [55]▲	5.81	4.96	6.23	0.83
SynRhythm [44]◆	10.88	-	11.08	-
RhythmNet [45]◆	3.99	-	3.99	0.87
HR-CNN [53]★	-	7.25	9.24	0.51
rPPGNet [67]★	7.82	5.51	7.82	0.78
DeepPhys [11]★	-	4.57	-	-
AutoHR [63]★	4.73	3.78	5.10	<u>0.86</u>
Meta-rPPG [28]★	4.9	3.01	3.68	0.85
PhysFormer (Ours)★	3.87	<u>3.25</u>	<u>3.97</u>	0.87

HR estimation on MAHNOB-HCI. For the HR estimation tasks on MAHNOB-HCI, similar to [67], subject-independent 9-fold cross-validation protocol is adopted. In consideration of the convergence difficulty due to the low illumination and high compression videos in MAHNOB-HCI, we finetune the VIPL-HR pretrained model on MAHNOB-HCI for further 15 epochs. The HR estimation results are shown in Table 3. The proposed PhysFormer achieves the lowest SD (3.87 bpm) and highest *r* (0.87) among the traditional, non-end-to-end learning, and end-to-end learning methods, which indicates the reliability of the learned rPPG features from PhysFormer under sufficient supervision. Our performance is on par with the latest end-to-end learning method Meta-rPPG [28] without transductive adaptation from target frames.

HR, HRV and RF estimation on OBF. We also conduct experiments for three types of physiological signals, i.e., HR, RF, and HRV measurement on the OBF [29] dataset. Following [46, 67], we use a 10-fold subject-exclusive protocol for all experiments. All the results are shown in Table 2. From the results, we can see that the proposed approach outperforms the existing state-of-the-art traditional (ROI_green [29], CHROM [13], POS [59]) and end-to-end learning (rPPGNet [67]) methods by a large margin on all evaluation metrics for HR, RF and all HRV features. The proposed PhysFormer also gives more accurate estimation

Table 4. Cross-dataset results on MMSE-HR [55].

Method	SD ↓ (bpm)	MAE ↓ (bpm)	RMSE ↓ (bpm)	<i>r</i> ↑
Li2014 [30]▲	20.02	-	19.95	0.38
CHROM [13]▲	14.08	-	13.97	0.55
Tulyakov2016 [55]▲	12.24	-	11.37	0.71
ST-Attention [47]◆	9.66	-	10.10	0.64
RhythmNet [45]◆	6.98	-	7.33	0.78
CVD [46]◆	6.06	-	6.04	0.84
PhysNet [65]★	12.76	-	13.25	0.44
TS-CAN [34]★	-	3.85	7.21	0.86
AutoHR [63]★	<u>5.71</u>	-	5.87	<u>0.89</u>
EfficientPhys-C [35]★	-	<u>2.91</u>	<u>5.43</u>	0.92
EfficientPhys-T1 [35]★	-	3.48	7.21	0.86
PhysFormer (Ours)★	5.22	2.84	5.36	0.92

in terms of HR, RF, and LF/HF compared with the preprocessed signal map based non-end-to-end learning method CVD [46]. These results indicate that PhysFormer could not only handle the average HR estimation task but also give a promising prediction of the rPPG signal for RF measurement and HRV analysis, which shows its potential in many healthcare applications.

4.4. Cross-dataset Testing

Besides of the intra-dataset testings on the VIPL-HR, MAHNOB-HCI, and OBF datasets, we also conduct cross-dataset testing on MMSE-HR [55] following the protocol of [45]. The models trained on VIPL-HR are directly tested on MMSE-HR. All the results of the proposed approach and the state-of-the-art methods are shown in Table 4. It is clear that the proposed PhysFormer generalizes well in unseen domain. It is worth noting that PhysFormer achieves the lowest SD (5.22 bpm), MAE (2.84 bpm), RMSE (5.36 bpm) as well as the highest *r* (0.92) among the traditional, non-end-to-end learning and end-to-end learning based methods, indicating 1) the predicted HRs are highly correlated with the ground truth HRs, and 2) the model learns domain-invariant intrinsic rPPG-aware features. Compared with the spatio-temporal transformer based EfficientPhys-T1 [35], our proposed PhysFormer is able to predict more accurate physiological signals, which indicates the effectiveness of the long-range spatio-temporal attention.

Table 5. Ablation of Tube Tokenization of PhysFormer. The three dimensions in tensors indicate length \times height \times width.

Inputs	[Stem] Feature Size	[Tube Size] Token Numbers	RMSE ↓ (bpm)
$160 \times 128 \times 128$	$[\times]$ $160 \times 128 \times 128$	$[4 \times 32 \times 32]$ $40 \times 4 \times 4$	10.62
$160 \times 128 \times 128$	$[\sqrt{ }]$ $160 \times 16 \times 16$	$[4 \times 4 \times 4]$ $40 \times 4 \times 4$	7.56
$160 \times 96 \times 96$	$[\sqrt{ }]$ $160 \times 12 \times 12$	$[4 \times 4 \times 4]$ $40 \times 3 \times 3$	8.03
$160 \times 128 \times 128$	$[\sqrt{ }]$ $160 \times 16 \times 16$	$[4 \times 16 \times 16]$ $40 \times 1 \times 1$	10.61
$160 \times 128 \times 128$	$[\sqrt{ }]$ $160 \times 16 \times 16$	$[2 \times 4 \times 4]$ $80 \times 4 \times 4$	7.81

Table 6. Ablation of TD-MHSA and ST-FF in PhysFormer.

MHSA	τ	Feed-forward	RMSE (bpm) ↓
-	-	ST-FF	9.81
TD-MHSA	$\sqrt{D_h} \approx 4.9$	ST-FF	9.51
TD-MHSA	2.0	ST-FF	7.56
vanilla MHSA	2.0	ST-FF	10.43
TD-MHSA	2.0	vanilla FF	8.27

Table 7. Ablation of dynamic loss in the frequency domain. The temporal loss \mathcal{L}_{time} is with fixed $\alpha=0.1$ here. ‘CE’ and ‘LD’ denote cross-entropy and label distribution, respectively.

Frequency loss	β	Strategy	RMSE (bpm) ↓
$\mathcal{L}_{CE} + \mathcal{L}_{LD}$	1.0	fixed	8.48
$\mathcal{L}_{CE} + \mathcal{L}_{LD}$	5.0	fixed	8.86
$\mathcal{L}_{CE} + \mathcal{L}_{LD}$	[1.0, 5.0]	linear	8.37
$\mathcal{L}_{CE} + \mathcal{L}_{LD}$	[1.0, 5.0]	exponential	7.56
\mathcal{L}_{CE}	[1.0, 5.0]	exponential	8.09
\mathcal{L}_{LD}	[1.0, 5.0]	exponential	8.21
\mathcal{L}_{LD} (real distribution)	[1.0, 5.0]	exponential	8.72

4.5. Ablation Study

We also provide the results of ablation studies for HR estimation on the **Fold-1** of the VIPL-HR [45] dataset.

Impact of tube tokenization. In the default setting of PhysFormer, a shallow stem cascaded with a tube tokenization is used. In this ablation, we consider other four tokenization configurations with or w/o stem. It can be seen from the first row in Table 5 that the stem helps the PhysFormer see better [61], and the RMSE increases dramatically (+3.06 bpm) when w/o the stem. Then we investigate the impacts of the spatial and temporal domains in tube tokenization. It is clear that the result in the fourth row with full spatial projection is quite poor (RMSE=10.61 bpm), indicating the necessity of the spatial attention. In contrast, tokenization with smaller tempos (e.g., [2x4x4]) or spatial inputs (e.g., 160x96x96) reduces performance slightly.

Impact of TD-MHSA and ST-FF. As shown in Table 6, both the TD-MHSA and ST-FF play vital roles in PhysFormer. The result in the first row shows that the performance degrades sharply without spatio-temporal attention. Moreover, it can be seen from the last two rows

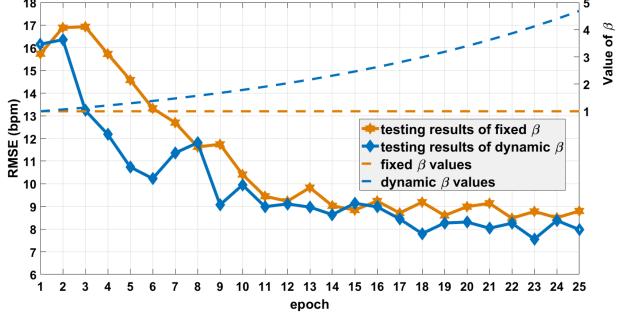


Figure 3. Testing results of fixed and dynamic frequency supervisions on the Fold-1 of VIPL-HR.

that without TD-MHSA/ST-FF, PhysFormer with vanilla MHSA/FF obtains 10.43/8.27 bpm RMSE. One important finding in this research is that, the temperature τ influences the MHSA a lot. When the $\tau = \sqrt{D_h}$ like previous ViT [1, 15], the predicted rPPG signals are unsatisfied (RMSE=9.51 bpm). Regularizing the τ with smaller value enforces sparser spatio-temporal attention, which is effective for the quasi-periodic rPPG task.

Impact of label distribution learning. Besides the temporal loss \mathcal{L}_{time} and frequency cross-entropy loss \mathcal{L}_{CE} , the ablations w/ and w/o label distribution loss \mathcal{L}_{LD} are shown in the last four rows of Table 7. Although the \mathcal{L}_{LD} performs slightly worse (+0.12 bpm RMSE) than \mathcal{L}_{CE} , the best performance can be achieved using both losses, indicating the effectiveness of explicit distribution constraints for extreme-frequency interference alleviation and adjacent label knowledge propagation. It is interesting to find from the last two rows that using real PSD distribution from groundtruth PPG signals as p , the performance is inferior due to the lack of an obvious peak and partial noise. We can also find from the Fig. 4(a) that the σ ranged from 0.9 to 1.2 for \mathcal{L}_{LD} are suitable to achieve good performance.

Impact of dynamic supervision. Fig. 3 illustrates the testing performance on Fold-1 VIPL-HR when training with fixed and dynamic supervision. It is clear that with exponential increased frequency loss, models in the blue curve converge faster and achieve smaller RMSE. We also compare several kinds of fixed and dynamic strategies in Table 7. The results in the first four rows indicate 1) using fixed higher β leads to poorer performance caused by the convergency difficulty; 2) models with the exponentially increased β perform better than using linear increment.

Impact of θ and layer/head numbers. Hyperparameter θ tradeoffs the contribution of local temporal gradient information. As illustrated in Fig. 3(b), PhysFormer could achieve smaller RMSE when $\theta=0.4$ and 0.7, indicating the importance of the normalized local temporal difference features for global spatio-temporal attention. We also investigate how the layer and head numbers influence the performance. As shown in Fig. 5(a), with deeper temporal trans-

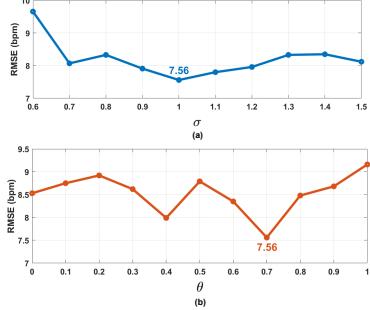


Figure 4. Impacts of the (a) σ in label distribution learning and (b) θ in TD-MHSA.

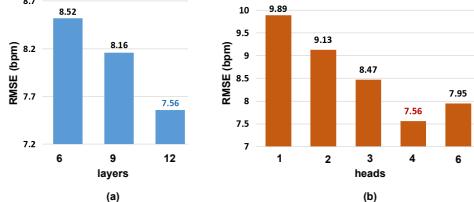


Figure 5. Ablation of the (a) layers and (b) heads in PhysFormer.

former blocks, the RMSE are reduced progressively despite heavier computational cost. In terms of the impact of head numbers, it is clear to find from Fig. 5(b) that PhysFormer with four heads perform the best while fewer heads lead to sharp performance drops.

4.6. Visualization and Discussion

We visualize the attention map from the last TD-MHSA module as well as one example about the query-key interaction in Fig. 6. The x and y axes indicate the attention confidence from key and query tube tokens, respectively. From the attention map, we can easily find periodic or quasi-periodic responses along both axes, indicating the periodicity of the intrinsic rPPG features from PhysFormer. To be specific, given the 530th tube token (in blue) from the forehead (spatial face domain) and peak (temporal signal domain) locations as a query, the corresponding key responses are illustrated at the blue line in the attention map. On one hand, it can be seen from the key responses that dominant spatial attentions focus on the facial skin regions and discard unrelated background. On the other hand, the temporal localizations of the key responses are around peak positions in the predicted rPPG signals. All these patterns are reasonable: 1) the forehead and cheek regions [57] have richer blood volume for rPPG measurement and are also reliable since these regions are less affected by facial muscle movements due to e.g., facial expressions, talking; and 2) rPPG signals from healthy people are usually periodic.

However, we also find two limitations of the spatio-temporal attention from Fig. 6. First, there are still some unexpected responses (e.g., continuous query tokens with similar key responses) in the attention map, which might introduce task-irrelevant noise and damage to the perfor-

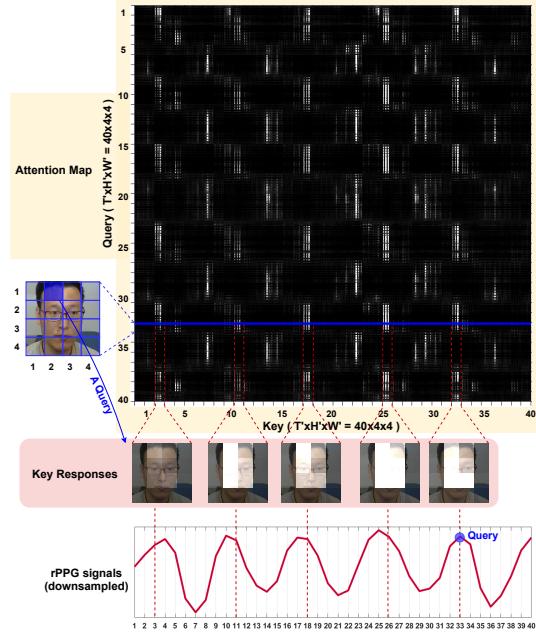


Figure 6. Visualization of the attention map from the 1st head in last TD-MHSA module. Given the 530th tube token in blue as a query, representative key responses are illustrated (the brighter, the more attentive). The predicted downsampled rPPG signals are shown for temporal attention understanding.

mance. Second, the temporal attentions are not always accurate, and some are coarse with phase shifts (e.g., the first vertical dotted line of the rPPG signals in bottom Fig. 6).

5. Conclusions and Future Work

In this paper, we propose an end-to-end video transformer architecture, namely PhysFormer, for remote physiological measurement. With temporal difference transformer and elaborate supervisions, PhysFormer is able to achieve superior performance on benchmark datasets. The study of video transformer based physiological measurement is still at an early stage. Future directions include: 1) Designing more efficient architectures. The proposed PhysFormer is with 7.03 M parameters and 47.01 GFLOPs, which is unfriendly for mobile deployment; 2) Exploring more accurate yet efficient spatio-temporal self-attention mechanism especially for long sequence rPPG monitoring.

Acknowledgment This work was supported by the Academy of Finland for Academy Professor project EmotionAI (grants 336116, 345122), and ICT 2023 project (grant 328115), by Ministry of Education and Culture of Finland for AI forum project, the National Natural Science Foundation of China (Grant No. 62002283), the EP-SRC grant: Turing AI Fellowship: EP/W002981/1, EP-SRC/MURI grant EP/N019474/1. We would also like to thank the Royal Academy of Engineering and FiveAI. As well, the authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

References

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *ICCV*, 2021.
- [2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009.
- [3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv:2102.05095*, 2021.
- [4] Adrian Bulat, Juan-Manuel Perez-Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. *NeurIPS*, 2021.
- [5] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv:2106.06847*, 2021.
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.
- [7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.
- [8] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *ICCV*, 2021.
- [9] Haoyu Chen, Hao Tang, Nicu Sebe, and Guoying Zhao. Anifomer: Data-driven 3d animation with transformer. *BMVC*, 2021.
- [10] Haoyu Chen, Hao Tang, Zitong Yu, Nicu Sebe, and Guoying Zhao. Geometry-contrastive transformer for generalized 3d pose transfer. In *AAAI*, 2022.
- [11] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *ECCV*, 2018.
- [12] Xun Chen, Juan Cheng, Rencheng Song, Yu Liu, Rabab Ward, and Z Jane Wang. Video-based heart rate measurement: Recent advances and future prospects. *IEEE Transactions on Instrumentation and Measurement*, 2018.
- [13] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 2013.
- [14] Mingyu Ding, Xiaochen Lian, Linjie Yang, Peng Wang, Xiaojie Jin, Zhiwu Lu, and Ping Luo. Hr-nas: Searching efficient high-resolution neural architectures with lightweight transformers. In *CVPR*, 2021.
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.
- [16] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *ICCV*, 2021.
- [17] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE TIP*, 2017.
- [18] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *IJCAI*, 2018.
- [19] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE TPAMI*, 2013.
- [20] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *ICCV*, 2021.
- [21] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, 2019.
- [22] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv:2012.12556*, 2020.
- [23] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv:2103.00112*, 2021.
- [24] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. *ICCV*, 2021.
- [25] Gee-Sern Hsu, ArulMurugan Ambikapathi, and Ming-Shiang Chen. Deep learning with time-frequency representation for pulse estimation from facial videos. In *IJCB*, 2017.
- [26] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv:2101.01169*, 2021.
- [27] Antony Lam and Yoshinori Kuno. Robust heart rate measurement from video using select random patches. In *ICCV*, 2015.
- [28] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *ECCV*, 2020.
- [29] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapiro Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *FG*, 2018.
- [30] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *CVPR*, 2014.
- [31] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *CVPR*, 2019.
- [32] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *arXiv:2106.04554*, 2021.
- [33] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *ICCV*, 2021.
- [34] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *NeurIPS*, 2020.
- [35] Xin Liu, Brian L Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based vitals measurement. *arXiv:2110.04447*, 2021.

[36] Xin Liu, Shwetak Patel, and Daniel McDuff. Camera-based physiological sensing: Challenges and future directions. *arXiv:2110.13362*, 2021.

[37] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *arXiv:2106.10271*, 2021.

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*, 2021.

[39] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv:2106.13230*, 2021.

[40] Hao Lu and Hu Han. Nas-hr: Neural architecture search for heart rate estimation from face videos. *Virtual Reality & Intelligent Hardware*, 2021.

[41] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *CVPR*, 2021.

[42] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv:2102.00719*, 2021.

[43] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Continuous heart rate measurement from face: A robust rppg approach with distribution learning. In *IJCB*, 2017.

[44] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a deep heart rate estimator from general to specific. In *ICPR*, 2018.

[45] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE TIP*, 2019.

[46] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *ECCV*, pages 295–310. Springer, 2020.

[47] Xuesong Niu, Xingyuan Zhao, Hu Han, Abhijit Das, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Robust remote heart rate estimation from face utilizing spatial-temporal attention. In *FG*, 2019.

[48] Ewa M Nowara, Daniel McDuff, and Ashok Veeraraghavan. The benefit of distraction: Denoising camera-based physiological measurements using inverse attention. In *ICCV*, 2021.

[49] Ming-Zher Poh, Daniel McDuff, and Rosalind Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 2010.

[50] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 2010.

[51] Ying Qiu, Yang Liu, Juan Arteaga-Falconi, Haiwei Dong, and Abdulmotaleb El Saddik. Evm-cnn: Real-time contactless heart rate estimation from facial video. *IEEE TMM*, 2018.

[52] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 2011.

[53] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *BMVC*, 2018.

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

[55] 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 *CVPR*, 2016.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[57] Wim Verkruyse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 2008.

[58] Lining Wang, Haosen Yang, Wenhao Wu, Hongxun Yao, and Hujie Huang. Temporal action proposal generation with transformers. *arXiv:2105.12043*, 2021.

[59] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 2017.

[60] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *ICCV*, 2021.

[61] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *NeurIPS*, 2021.

[62] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhiuwen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *arXiv:2107.03377*, 2021.

[63] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. *IEEE SPL*, 2020.

[64] Zitong Yu, Xiaobai Li, Pichao Wang, and Guoying Zhao. Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection. *IEEE SPL*, 2021.

[65] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *BMVC*, 2019.

[66] Zitong Yu, Xiaobai Li, and Guoying Zhao. Facial-video-based physiological signal measurement: Recent advances and affective applications. *IEEE Signal Processing Magazine*, 2021.

[67] 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 *ICCV*, 2019.

[68] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: a survey. *arXiv:2106.14948*, 2021.

[69] Zitong Yu, Benjia Zhou, Jun Wan, Pichao Wang, Haoyu Chen, Xin Liu, Stan Z Li, and Guoying Zhao. Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. *IEEE TIP*, 2021.

- [70] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv:2101.11986*, 2021.
- [71] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, 2020.
- [72] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE SPL*, 2016.
- [73] Jiaojiao Zhao, Xinyu Li, Chunhui Liu, Shuai Bing, Hao Chen, Cees GM Snoek, and Joseph Tighe. Tuber: Tube-transformer for action detection. *arXiv:2104.00969*, 2021.
- [74] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021.