ACuRE: Accurate Continuity-Regularized SpO_2 Estimation Using Liquid Time-Constant Networks: Supplementary

Shahzad Ahmad[1] shahzaa@hiof.no
Divya Mishra[3] divya.mishra@ddn.upes.ac.in
Sania Bano[2] sania.22eez0012@iitrpr.ac.in
Sukalpa Chanda [1] sukalpa@ieee.org
Yogesh Singh Rawat[4] yogesh@crcv.ucf.edu
[1]Østfold University College, Norway [2]Indian Institute of Technology Ropar, India
[3]University of Petroleum and Energy Studies,India [4]University of Central Florida, USA

Contents

1. Comparison of Temporal Models

To evaluate the effectiveness of the Liquid Time-Constant (LTC) network in our ACuRE framework for non-contact SpO2 estimation, we conducted an ablation study comparing LTC against several alternative temporal modeling approaches: Neural Ordinary Differential Equations (N-ODE) [4], Continuous-Time Recurrent Neural Networks (CTRNN) [8], Long Short-Term Memory (LSTM) [12], Gated Recurrent Unit (GRU) [6], vanilla Recurrent Neural Network (RNN) [18], and Temporal Convolutional Network (TCN) [2]. These models were integrated into the ACuRE pipeline in place of the LTC module, keeping the two-branch 3D-ResNet-18 for AC/DC signal separation and the physics-informed PDE loss unchanged. Performance was evaluated on three datasets: PURE [19], BH-rPPG [24], and VIPL-HR [14], using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Pearson correlation coefficient (r) as metrics.

Table 1 presents the results. Across all datasets, ACuRE with LTC consistently outperforms alternative temporal models, achieving the lowest MAE and RMSE and the highest correlation. For instance, on the PURE dataset, LTC achieves an MAE of 0.28 ± 0.17, compared to 0.50 ± 0.22 for LSTM and 3.01 ± 1.36 for TCN. The VIPL-HR dataset, which is more challenging due to variations in illumination and motion, shows a similar trend, with LTC achieving an MAE of 0.72 ± 0.20, significantly better than TCN's 5.19 ± 1.85. N-ODE and CTRNN perform competitively but fall short of LTC, particularly on VIPL-HR, where CTRNN's MAE is 1.49 ± 0.66. GRU and RNN, while outperforming TCN, exhibit higher errors than LSTM (e.g., GRU: 0.52 ± 0.23 on PURE; RNN: 1.10 ± 0.32 on VIPL-HR), likely due to their simpler architectures struggling with the continuous dynamics of physiological signals.

The superior performance of LTC can be attributed to its ability to model continuous-time dynamics, which aligns well with the temporal continuity of photoplethysmography signals, as noted in prior work [9]. In contrast, discrete-time models like LSTM and GRU struggle with long-term dependencies, while TCN's convolutional approach is less suited for capturing the smooth, periodic nature of rPPG signals. These results validate the choice of LTC in ACuRE and highlight its robustness across diverse datasets.

2. Cross-Dataset Evaluation

This table shows cross-dataset evaluation between **PURE** (30 fps), **VIPL-HR** (25 fps), and **BH-rPPG** (15 fps), where we train on one dataset and test on a different one. The main observation is that transfers involving *BH-rPPG* are the hardest: at 15 fps the pulse signal is sampled less often, making it more prone to distortion/aliasing, and BH-rPPG also has stronger lighting and motion changes. Transfers with VIPL-HR are milder, and models trained on VIPL-HR often generalize better to PURE because VIPL-HR covers a wider range of subjects, devices, and motion. ACuRE handles these shifts better by (i) separating steady color/light-

Table 1. Performance comparison of temporal models in the ACuRE framework across PURE, BH-rPPG, and VIPL-HR datasets. Metrics are reported as mean ± standard deviation. Best results are in **bold**.

Model	MAE	RMSE	r
PURE Dataset			
ACuRE (LTC)	**0.28 ± 0.17**	**0.50 ± 0.16**	**0.95 ± 0.03**
N-ODE	0.33 ± 0.14	0.58 ± 0.12	0.93 ± 0.03
CTRNN	0.36 ± 0.11	0.60 ± 0.10	0.92 ± 0.08
LSTM	0.50 ± 0.22	0.70 ± 0.20	0.85 ± 0.05
GRU	0.52 ± 0.23	0.72 ± 0.21	0.84 ± 0.06
RNN	0.60 ± 0.25	0.80 ± 0.23	0.80 ± 0.07
TCN	3.01 ± 1.36	1.70 ± 0.35	0.29 ± 0.35
BH-rPPG Dataset			
ACuRE (LTC)	**0.39 ± 0.18**	**0.61 ± 0.13**	**0.92 ± 0.04**
N-ODE	0.42 ± 0.12	0.63 ± 0.15	0.91 ± 0.02
CTRNN	0.46 ± 0.16	0.68 ± 0.16	0.90 ± 0.02
LSTM	0.60 ± 0.25	0.82 ± 0.22	0.80 ± 0.06
GRU	0.62 ± 0.26	0.84 ± 0.23	0.79 ± 0.06
RNN	0.70 ± 0.28	0.92 ± 0.25	0.75 ± 0.07
TCN	2.93 ± 0.64	1.70 ± 0.18	0.49 ± 0.04
VIPL-HR Dataset			
ACuRE (LTC)	**0.72 ± 0.20**	**0.84 ± 0.12**	**0.87 ± 0.04**
N-ODE	0.79 ± 0.38	0.80 ± 0.21	0.86 ± 0.03
CTRNN	1.49 ± 0.66	1.19 ± 0.27	0.72 ± 0.18
LSTM	0.90 ± 0.28	1.05 ± 0.25	0.78 ± 0.07
GRU	0.93 ± 0.29	1.08 ± 0.26	0.77 ± 0.08
RNN	1.10 ± 0.32	1.20 ± 0.28	0.70 ± 0.09
TCN	5.19 ± 1.85	2.25 ± 0.37	0.40 ± 0.16

ing (DC) from pulsatile changes (AC) so illumination differences do not contaminate the signal, (ii) using a continuous-time temporal module that is less tied to a single frame rate, and (iii) adding a simple physics-based continuity regularizer that keeps the predicted signal smooth and consistent. Together, these choices make ACuRE less sensitive to frame-rate and appearance differences across datasets, explaining the consistent gains in Table 2.

3. Backbone Ablation Study

To further assess the adaptability of the proposed ACuRE framework, we evaluated it with different backbone architectures for non-contact SpO_2 estimation. Specifically, we considered ResNet3D18 (the backbone used in the main system), PhysNet [23], ResNet2D18 [11], a 2D Vision Transformer (ViT2D) [7], and a 3D Vision Transformer (ViT3D) [3]. These architectures span convolutional and transformer-based paradigms in both 2D and 3D settings, enabling a broad comparison of modeling choices.

Experiments were conducted on the PURE, VIPL-HR, and BH-rPPG datasets, with performance evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Pearson correlation coefficient (r). Results are summarized in Table 3.

The findings highlight the flexibility of ACuRE's design: it maintains strong performance across diverse backbones due to its physics-informed pipeline, which includes the PDE loss and the Liquid Time-Constant (LTC) module. These components provide temporal and physiological consistency, complementing a wide range of feature extractors. While alternative backbones (e.g., PhysNet or ViT-based models) achieve competitive results in relatively controlled datasets such as PURE or BH-rPPG, ResNet3D18 consistently delivers the most robust performance, particularly on VIPL-HR. We attribute this to the larger scale and greater variability of VIPL-HR, which includes diverse motion, illumination, and device conditions, where 3D convolutional features better capture spatio-temporal variations.

3.1. Justification for Selecting ResNet3D18

We selected ResNet3D18 as the primary backbone in the main paper due to its superior robustness and performance across diverse datasets, particularly in challenging real-world conditions. As shown in Table 3, ResNet3D18 achieves the lowest MAE (0.72), RMSE (0.84), and highest correlation (0.87) in the VIPL-HR dataset, outperforming alternatives by a significant margin (e.g., 30% lower MAE than ViT3D and over 50% lower than PhysNet and ViT2D). This dataset's challenging conditions, including noise, illumination variations, and motion artifacts, highlight ResNet3D18's ability to generalize effectively for rPPG applications [16]. In contrast, while PhysNet excels in controlled settings like PURE (MAE: 0.21), it struggles in VIPL-HR (MAE: 1.42), indicating limited robustness. Similarly, ViT2D and ViT3D perform well in PURE and BH-rPPG (e.g., ViT2D's MAE of 0.30 and r of 0.93 in BH-rPPG) but exhibit higher errors and variance in VIPL-HR, suggesting sensitivity to noise.

The 3D convolutional architecture of ResNet3D18 effectively captures spatial-temporal dynamics essential for AC/DC signal separation in rPPG tasks [21], significantly outperforming its 2D counterpart (ResNet2D18) across all metrics and datasets (e.g., 60% lower MAE in VIPL-HR). This justifies the use of 3D modeling over 2D variants, aligning with prior work on spatiotemporal networks for physiological signal estimation [16]. Furthermore, ResNet3D18's consistent high correlation ($r \geq 0.87$ across datasets) ensures reliable SpO_2 predictions, making it the optimal choice for the ACuRE framework. The strong performance of PhysNet, ViT2D, and ViT3D in PURE and BH-rPPG underscores the ACuRE framework's flexibility, as its physics-informed design and LTC module en-

Table 2. **Cross-dataset SpO_2 generalization (no target-domain fine-tuning).** Train on one dataset, test on another. ↑: higher is better; ↓: lower is better. Datasets have different frame rates: PURE (30 fps), VIPL-HR (25 fps), BH-rPPG (15 fps). Pairs involving BH-rPPG are more challenging due to the 15 fps temporal mismatch.

Train	Test	Method	MAE↓	r↑
PURE (30 fps)	BH-rPPG (15 fps)	3D-CNN+LSTM	**2.5 ± 0.9**	**0.50 ± 0.05**
		Akamatsu et al. [1]	2.45 ± 0.30	0.55 ± 0.05
		ACuRE (ours)	**2.28 ± 0.27**	**0.60 ± 0.06**
	VIPL-HR (25 fps)	**ACuRE (ours)**	2.22 ± 0.47	0.57 ± 0.02
BH-rPPG (15 fps)	PURE (30 fps)	3D-CNN+LSTM	2.29 ± 0.34	0.50 ± 0.06
		Akamatsu et al. [1]	2.05 ± 0.28	0.57 ± 0.03
		ACuRE (ours)	**1.86 ± 0.24**	**0.60 ± 0.02**
	VIPL-HR (25 fps)	**ACuRE (ours)**	2.54 ± 0.24	0.50 ± 0.02
VIPL-HR (25 fps)	PURE (30 fps)	3D-CNN+LSTM	1.11 ± 0.17	0.75 ± 0.09
	PURE (30 fps)	**ACuRE (ours)**	**0.83 ± 0.11**	**0.85 ± 0.10**
	BH-rPPG (15 fps)	**ACuRE (ours)**	1.58 ± 0.21	0.50 ± 0.03

Table 3. Backbone ablation study results for SpO_2 estimation in the ACuRE framework. Metrics are reported as mean ± standard deviation. Bold values indicate the best performance per metric and dataset.

Backbone	PURE			VIPL-HR			BH-rPPG		
	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r
ResNet3D18 (Ours)	0.28 ± 0.17	0.50 ± 0.16	**0.95 ± 0.03**	**0.72 ± 0.20**	**0.84 ± 0.12**	**0.87 ± 0.04**	0.39 ± 0.18	0.61 ± 0.13	**0.92 ± 0.04**
PhysNet [23]	**0.21 ± 0.11**	**0.44 ± 0.11**	0.88 ± 0.09	1.42 ± 0.54	1.17 ± 0.21	0.74 ± 0.08	0.44 ± 0.19	0.65 ± 0.15	0.90 ± 0.04
ResNet2D18 [11]	0.45 ± 0.09	0.67 ± 0.07	0.69 ± 0.25	2.83 ± 1.53	1.63 ± 0.43	0.48 ± 0.18	1.81 ± 0.84	1.31 ± 0.32	0.56 ± 0.22
ViT2D [7]	**0.22 ± 0.13**	0.46 ± 0.12	0.86 ± 0.12	1.42 ± 0.84	1.15 ± 0.31	0.65 ± 0.27	**0.30 ± 0.10**	**0.54 ± 0.09**	**0.93 ± 0.02**
ViT3D [3]	**0.21 ± 0.08**	0.45 ± 0.08	0.88 ± 0.08	1.03 ± 0.35	0.99 ± 0.20	0.77 ± 0.16	0.37 ± 0.16	0.59 ± 0.12	**0.91 ± 0.02**

hance performance across diverse architectures. However, ResNet3D18's superior generalization in noisy, real-world-like conditions (VIPL-HR) makes it the preferred choice for robust SpO_2 estimation, supporting its adoption in the main paper.

4. Complexity and Efficiency

What we have measure. For each temporal module (LTC, RNN, GRU, LSTM, CTRNN, Neural ODE, TCN) plugged into the same AC/DC twin 3D-ResNet-18 backbone, we report: (i) parameter counts with a component breakdown (Backbone / Head / Temporal / Loss), (ii) forward FLOPs/clip (fvcore), (iii) *test-time inference latency* (ms/batch), and (iv) training time *with* and *without* the PDE continuity loss (LPDE).[1] A full per-component summary is provided in Table 4.

How we have computed Complexity and Efficiency. FLOPs are obtained with `fvcore` on a float32 copy of the model and a single-clip input (to avoid AMP/dtype issues). Timings use a unified PyTorch script with automatic mixed precision (`autocast+GradScaler`) and synchronized measurements over warm-up & repeated trials. All variants share the same input: batch $B=2$, $T=300$ frames, 32×32 pixels. The exact numbers reported in Table 4 come from this setup.

Why intra-backbone comparisons. Absolute wall-clock across different papers is not directly comparable due to differing backbones, input sizes, training loops, and hardware. To be fair and reproducible, we keep the backbone, input resolution, and code path fixed, and vary only the temporal block. This isolates the incremental cost/benefit of each temporal design under identical conditions, making the deltas in Table 4 directly interpretable. We also report implementation-agnostic metrics (FLOPs/params) that are stable across hardware.

PDE loss details. LPDE regularizes the AC stream $\hat{\mathbf{V}}_{AC} = [\rho, u, v] \in \mathbb{R}^{3\times T\times H\times W}$ via the discrete continuity residual $R = \partial_t\rho + \partial_x(\rho u) + \partial_y(\rho v)$ implemented with finite differences. LPDE adds *zero parameters*, is evaluated *only during training*, and consists of simple elementwise ops ($\mathcal{O}(BTHW)$). Consequently, its overhead is small in practice, as quantified in Table 4.

Why is LTC relatively slower? Although LTC adds very few FLOPs (0.018 G) and parameters (0.27 M), its update is unrolled for L steps with per-step *dynamic* time-constant computation and several elementwise operations (tanh/sigmoid/mul/div). In PyTorch these appear as many

[1] FLOPs are forward-only; training overhead of LPDE is captured via wall-clock timing. `fvcore` may not symbolically count some pointwise ops; this has negligible effect because 3D convolutions dominate.

Temporal	Params (M)				FLOPs (G/clip)				Infer (ms/b)	Training time (ms/batch)		
	Total	Back	Head	Temp	Back	Head	Temp	Total		Base	+LPDE	Δ (%)
LTC	66.91	66.33	0.31	0.27	125.04	0.0003	0.0181	125.06	44.1	165.4	167.0	+1.57 (**0.9**)
RNN	67.00	66.33	0.31	0.36	125.04	0.0003	0.0002	125.05	17.0	69.7	69.9	+0.21 (0.3)
GRU	67.36	66.33	0.31	0.72	125.04	0.0003	0.0002	125.05	17.5	70.4	70.8	+0.38 (0.5)
LSTM	67.54	66.33	0.31	0.90	125.04	0.0003	0.0002	125.05	17.5	70.5	71.1	+0.56 (0.8)
CTRNN	66.91	66.33	0.31	0.27	125.04	0.0003	0.0181	125.06	27.3	120.2	121.3	+1.07 (0.9)
NODE	66.89	66.33	0.31	0.25	125.04	0.0003	0.0631	125.11	106.3	324.5	325.7	+1.29 (0.4)
TCN	69.88	66.33	0.31	3.24	125.04	0.0003	0.3920	125.44	17.6	71.3	72.1	+0.82 (**1.1**)

Table 4. **Complexity of ACuRE under a fixed backbone and input.** Backbone compute dominates (∼125 GFLOPs/clip); temporal heads add negligible FLOPs. LPDE adds *no* parameters and incurs only **0.3–1.1%** training overhead; inference is unaffected by LPDE.

small pointwise kernels, which reduces GPU fusion/tensor-core utilization. In contrast, RNN/GRU/LSTM use highly optimized cuDNN fused kernels, and TCN parallelizes time with 1D convolutions. This explains the modestly higher LTC latency despite comparable FLOPs.

Takeaways. (i) The AC/DC 3D-ResNet-18 backbone accounts for >99% of forward FLOPs, so ACuRE's temporal choice or the PDE term does not materially affect inference cost (Table 4). (ii) LPDE is training-only and adds ≤1.1% overhead in our controlled setting (Table 4). (iii) LTC trades a slightly higher latency (due to unrolled continuous-time updates with dynamic τ and pointwise ops) for a compact parameter/FLOP footprint (0.27 M / 0.018 G), while RNN/GRU/LSTM/TCN benefit from cuDNN/conv parallelism and thus run ∼17–18 ms/batch at test time for B=2, T=300, 32×32 (Table 4).

External complexity reports. We summarize compute/latency figures that prior SpO_2 papers explicitly report (when available). These numbers are *not* strictly comparable across works because input sizes, backbones, pre/post-processing, and hardware differ; hence we also provide our own controlled measurements in Sec. S3.

5. Theoretical Results

This appendix presents theorems and proofs supporting the Liquid Time-Constant (LTC) dynamics and continuity-regularized loss ($\mathcal{L}_{\text{PDE}}$) in ACuRE, as detailed in Section III. These results align with the LTC framework in [10] and our implementation (e.g., $\tau \in [0.1, 1.0]$, $\Delta t = 0.01$), extending beyond SpO_2 estimation to general video-based physiological signal modeling. Each theorem includes notation definitions and significance to connect strongly with ACuRE's contributions.

5.1. Stability of LTC Dynamics

Theorem 1. *For the LTC update*

$$x(t+\Delta t) = x(t) + \Delta t \frac{f(x(t), I(t), \theta) A}{1 + \Delta t \left(\frac{1}{\tau} + f(x(t), I(t), \theta)\right)}, \quad (1)$$

where $x(t) \in \mathbb{R}$ is the neuron state, $\tau \in [\tau_{\min}, \tau_{\max}]$ with $0 < \tau_{\min} \leq \tau_{\max} < \infty$ is the time constant, $A \in \mathbb{R}$ is the synaptic weight, $f(x, I, \theta) = W \tanh(\gamma I + \mu)$ is the activation function with $|f| \leq M = |W|$ (weights W, input scaling γ, bias μ), and $\Delta t > 0$ is the time step, if $\Delta t < \frac{\tau_{\min}}{1+M}$, the update is locally stable around the equilibrium $x^* = 0$ when $A = 0$.

Proof. Consider the continuous-time ODE governing ACuRE's LTC layer:

$$\frac{dx}{dt} = -\left[\frac{1}{\tau} + f(x, I, \theta)\right] x + f(x, I, \theta) A, \quad (2)$$

where $t \in \mathbb{R}_{\geq 0}$ is time, and $I(t) \in \mathbb{R}$ is the input signal (e.g., preprocessed video features from Section IV). To find equilibria, set:

$$0 = -\left[\frac{1}{\tau} + f(x^*, I, \theta)\right] x^* + f(x^*, I, \theta) A. \quad (3)$$

If $A = 0$ (no synaptic feedback, as in initial states):

$$-\left[\frac{1}{\tau} + f(x^*, I, \theta)\right] x^* = 0. \quad (4)$$

Since $\frac{1}{\tau} + f(x^*, I, \theta) > 0$ (with $\tau_{\min} = 0.1$, $f \geq -M$, and $M < 10$), $x^* = 0$ is an equilibrium. Linearize around $x^* = 0$:

$$\frac{dx}{dt} \approx -\left[\frac{1}{\tau} + f(0, I, \theta)\right] x. \quad (5)$$

The Jacobian is:

$$J = -\left[\frac{1}{\tau} + f(0, I, \theta)\right]. \quad (6)$$

Table 5. **Reported complexity/latency from prior video SpO_2 works (as available) and ACuRE.** Numbers are copied from the original papers; hardware, window length, and batch definitions differ, so these are not directly comparable. NR = not reported.

Method	Backbone	Params (M)	FLOPs (G)	Reported inference	Notes
MMFM [13]	RCA (CNN)	1.77	3.05	0.5 s / 50 frames	Table & latency note
CCM [13]	RCA (CNN)	1.08	2.05	0.5 s / 50 frames	Same setup
NBM [13]	RCA (CNN)	0.70	1.00	0.5 s / 50 frames	Same setup
CL-SPO2Net [17]	3D-CNN+CNN+BiLSTM	–	–	0.5 Hz (RTX 3080)	Fig. 7 caption
STMap+CNN [5]	EfficientNet-B3	9.2	1.0	–	Backbone-only complexity
STMap+CNN [5]	ResNet-50	26	4.1	–	Backbone-only complexity
STMap+CNN [5]	DenseNet-121	8.0	5.7	–	Backbone-only complexity
CCSpO2Net [20]	(foundation model)	–	–	NR	Paper emphasizes accuracy
ACuRE (ours)	twin 3D-ResNet-18 + LTC	**66.9**	**125.1**	**44.1 ms/batch**	B=2, 300 f, 32×32, RTX 8000

Bound J: since $\frac{1}{\tau} \geq \frac{1}{\tau_{\max}} = 1$ (per Section III.A) and $f(0, I, \theta) \in [-M, M]$,

$$J \leq -1 + M. \tag{7}$$

For $M < 1$ (e.g., $M = 0.5$ in practice), $J < 0$, ensuring stability of the ODE.

For the discrete update (Section III.A, implemented in `LTC.fused_step`):

$$g(x) = x + \Delta t \frac{f(x, I, \theta)A}{1 + \Delta t \left(\frac{1}{\tau} + f(x, I, \theta)\right)}, \tag{8}$$

evaluate at $x^* = 0$, $A = 0$:

$$g(0) = 0, \tag{9}$$

and compute the derivative:

$$g'(x) = 1 + \Delta t \frac{A\frac{\partial f}{\partial x}\left[1 + \Delta t\left(\frac{1}{\tau} + f\right)\right] - fA\Delta t\frac{\partial f}{\partial x}}{\left[1 + \Delta t\left(\frac{1}{\tau} + f\right)\right]^2}. \tag{10}$$

Since $f(x, I, \theta) = W \tanh(\gamma I + \mu)$ is independent of x (per `LTC.activation`), $\frac{\partial f}{\partial x} = 0$, simplifying to:

$$g'(0) = 1 - \Delta t \left(\frac{1}{\tau} + f(0, I, \theta)\right). \tag{11}$$

Stability requires $|g'(0)| < 1$:

$$-1 < 1 - \Delta t \left(\frac{1}{\tau} + f(0, I, \theta)\right) < 1. \tag{12}$$

Consider the lower bound: $\frac{1}{\tau} + f(0, I, \theta) \geq \frac{1}{\tau_{\min}} - M = 10 - M$. Then:

$$1 - \Delta t(10 - M) > -1 \implies \Delta t(10 - M) < 2, \tag{13}$$

$$\Delta t < \frac{2}{10 - M}. \tag{14}$$

The upper bound holds trivially (< 1). For $M = 0.5$, $\Delta t < \frac{2}{9.5} \approx 0.211$. The condition $\Delta t < \frac{\tau_{\min}}{1+M} = \frac{0.1}{1+0.5} = 0.066$ (Section III.A) is stricter and satisfied by $\Delta t = 0.01$, ensuring stability.

Significance: This stability ensures that LTC dynamics in ACuRE (Section III.A) remain bounded under video input perturbations (e.g., motion artifacts), critical for robust SpO_2 estimation, aligning with the bounded dynamics in [10]. □

5.2. Expressivity of LTC Dynamics

Theorem 2. *The LTC system approximates any continuously differentiable function $h(t) \in C^1([0, T], \mathbb{R})$ on a compact interval $[0, T]$ to an error $\epsilon > 0$, given sufficient neuron units $D \in \mathbb{N}$ and adjustable parameters $\tau \in [0.1, 1.0]$, $W, \gamma, \mu \in \mathbb{R}$, and $A \in \mathbb{R}$.*

Proof. The LTC ODE per `LTC.forward`) is given by:

$$\frac{dx}{dt} = -\left[\frac{1}{\tau} + f(x, I, \theta)\right] x + f(x, I, \theta)A, \tag{15}$$

where $x(t) \in \mathbb{R}$ is the state, $I(t) \in \mathbb{R}$ is the input, $\theta = (W, \gamma, \mu)$ parameterizes $f(x, I, \theta) = W \tanh(\gamma I + \mu)$, and $A \in \mathbb{R}$ is the synaptic weight. The goal is to approximate a target function $x(t) = h(t)$, satisfying:

$$\frac{dh}{dt} = -\left[\frac{1}{\tau} + f(h, I, \theta)\right] h + f(h, I, \theta)A. \tag{16}$$

Solve for the control term:

$$f(h, I, \theta)A = \frac{dh}{dt} + \left[\frac{1}{\tau} + f(h, I, \theta)\right] h. \tag{17}$$

Define:

$$u(t) = \frac{dh}{dt} + \frac{h}{\tau} + f(h, I, \theta)(h - A), \tag{18}$$

where $u(t)$ is continuous since $h \in C^1([0, T], \mathbb{R})$. Adjust the input $I(t) = I^*(t)$ such that:

$$f(h, I^*, \theta) \approx \frac{u(t)}{A}. \tag{19}$$

Since $f(x, I, \theta) = W \tanh(\gamma I + \mu)$ acts as a neural network over I (per `LTC.activation`), the universal approximation theorem [10] (Theorem 3) ensures that f can approximate $\frac{u(t)}{A}$ to within $\epsilon/2$ as $D \to \infty$ by tuning W, γ, μ. Consider the error dynamics:

$$\frac{d}{dt}(x - h) = -\left[\frac{1}{\tau} + f(x, I, \theta)\right] x + f(x, I, \theta)A - \frac{dh}{dt}. \tag{20}$$

Substitute $\frac{dh}{dt}$ from Equation (2):

$$\frac{d}{dt}(x - h) = -\left[\frac{1}{\tau} + f(x, I, \theta)\right](x - h) + [f(x, I, \theta) - f(h, I, \theta)](h - A). \tag{21}$$

With initial condition $x(0) = h(0)$, the solution is:

$$x(t) - h(t) = \int_0^t e^{-\int_s^t \left(\frac{1}{\tau} + f(x(u), I^*(u), \theta)\right) du} [f(x(s), I^*(s), \theta) - f(h(s), I^*(s), \theta)] (h(s) - A)\, ds. \tag{22}$$

Since $\frac{1}{\tau} + f > 0$ (as $\tau \geq 0.1$ and $|f| \leq |W| < 1$), the exponential term $e^{-\int_s^t \left(\frac{1}{\tau} + f\right) du} \leq e^{-(1-|W|)(t-s)}$ decays. Assume f is Lipschitz continuous in x with constant L_f (e.g., $L_f = |W| \cdot \max|\tanh'| < 1$), so:

$$|f(x(s), I^*, \theta) - f(h(s), I^*, \theta)| \leq L_f |x(s) - h(s)|. \tag{23}$$

If $f(x, I^*, \theta) \approx f(h, I^*, \theta)$ within $\epsilon/2$, bound the integrand:

$$|x(t) - h(t)| \leq \int_0^t e^{-(1-|W|)(t-s)} \left(\frac{\epsilon}{2}\right) |h(s) - A|\, ds. \tag{24}$$

Let $M = \sup_{s \in [0,T]} |h(s) - A| < \infty$ (since h is continuous on a compact interval and A is fixed). Then:

$$|x(t) - h(t)| \leq \frac{\epsilon}{2} M \int_0^t e^{-(1-|W|)(t-s)} \tag{25}$$

For $|W| < 1$, set $C = \frac{M}{1-|W|} < \infty$, so $|x(t) - h(t)| \leq C\frac{\epsilon}{2} < \epsilon$ by choosing ϵ small enough relative to C. Thus, the LTC approximates $h(t)$ within ϵ, completing the proof. □

Significance: This expressivity enables ACuRE to model complex SpO_2 signals (Section **??**.A), outperforming linear RNNs by capturing multi-scale dynamics, crucial for video-based physiological estimation.

5.3. Robustness of LTC Dynamics

Theorem 3. *For the LTC ODE in Equation (4), where $x(t) \in \mathbb{R}^D$ is the hidden state, $I(t) \in \mathbb{R}^D$ is the input, and $f(x, I, \theta) = W \tanh(\gamma I + \mu)$ is Lipschitz continuous with constant $L_f < \infty$, the dynamics satisfy:*

$$\left\| \frac{dx}{dt}(I + \delta I) - \frac{dx}{dt}(I) \right\| \leq L_f \|\delta I\|, \tag{26}$$

for any perturbation $\delta I \in \mathbb{R}^D$, ensuring bounded sensitivity to input variations.

Proof. Consider the LTC ODE from Section III (Equation 4):

$$\frac{dx}{dt} = -\left[\frac{1}{\tau} + f(x, I, \theta)\right] x + f(x, I, \theta)A. \tag{27}$$

For inputs $I(t)$ and perturbed $I(t) + \delta I(t)$, compute the difference:

$$\frac{dx}{dt}(I + \delta I) = -\left[\frac{1}{\tau} + f(x, I + \delta I, \theta)\right] x + f(x, I + \delta I, \theta)A, \tag{28}$$

$$\frac{dx}{dt}(I) = -\left[\frac{1}{\tau} + f(x, I, \theta)\right] x + f(x, I, \theta)A. \tag{29}$$

Subtract:

$$\frac{dx}{dt}(I + \delta I) - \frac{dx}{dt}(I) = \tag{30}$$

$$-[f(x, I + \delta I, \theta) - f(x, I, \theta)]\, x + \tag{31}$$

$$[f(x, I + \delta I, \theta) - f(x, I, \theta)]\, A. \tag{32}$$

Factorize:

$$\frac{dx}{dt}(I + \delta I) - \frac{dx}{dt}(I) = [f(x, I + \delta I, \theta) - f(x, I, \theta)]\,(A - x). \tag{33}$$

Since $f(x, I, \theta) = W \tanh(\gamma I + \mu)$ is Lipschitz continuous in I (as $\tanh$ has a bounded derivative ≤ 1), there exists $L_f < \infty$ such that:

$$\|f(x, I + \delta I, \theta) - f(x, I, \theta)\| \leq L_f \|\delta I\|. \tag{34}$$

Thus:

$$\left\| \frac{dx}{dt}(I + \delta I) - \frac{dx}{dt}(I) \right\| \leq \|f(x, I + \delta I, \theta) - f(x, I, \theta)\| \|A - x\|. \tag{35}$$

Bound the magnitude:

$$\left\| \frac{dx}{dt}(I + \delta I) - \frac{dx}{dt}(I) \right\| \leq L_f \|\delta I\| \|A - x\|. \tag{36}$$

Since $x(t)$ and A are finite in practice (per Section III.A and Theorem 1's stability), let $K = \sup_t \|A - x(t)\| < \infty$. Then:

$$\left\| \frac{dx}{dt}(I + \delta I) - \frac{dx}{dt}(I) \right\| \leq L_f K \|\delta I\|. \tag{37}$$

Defining $L_f' = L_f K$, we have:

$$\left\| \frac{dx}{dt}(I + \delta I) - \frac{dx}{dt}(I) \right\| \leq L_f' \|\delta I\|, \tag{38}$$

proving the robustness bound (Section III.A). □

5.4. Regularization Effect of Continuity Loss

Theorem 4. *Minimizing $\mathcal{L}_{PDE} = \frac{1}{TWH}\int R^2\,dt\,dx\,dy$, where $R = \frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{v})$ is the continuity residual, bounds the H^1-norm of ρ, reducing gradient variance, where $\rho(t,x,y) \in \mathbb{R}$ is the blood density field, and $\mathbf{v} = (v_x, v_y) \in \mathbb{R}^2$ is the velocity vector.*

Proof. Consider $\rho(t,x,y)$ on the domain $[0,T]\times[0,W]\times[0,H]$, representing the AC component's density channel (Section III-B), with periodic boundary conditions over time $t \in [0,T]$ and space $(x,y) \in [0,W]\times[0,H]$. The loss is:

$$\mathcal{L}_{\text{PDE}} = \frac{1}{TWH}\int\left(\frac{\partial\rho}{\partial t} + \frac{\partial(\rho v_x)}{\partial x} + \frac{\partial(\rho v_y)}{\partial y}\right)^2 dt\,dx\,dy, \tag{39}$$

where T, W, H are the temporal and spatial dimensions. Assume $|v_x|, |v_y| \le V < \infty$, bounding velocity magnitudes (e.g., blood flow in rPPG). The Sobolev H^1-norm is:

$$\|\rho\|_{H^1}^2 = \int\left(\rho^2 + \left|\frac{\partial\rho}{\partial t}\right|^2 + |\nabla\rho|^2\right) dt\,dx\,dy, \tag{40}$$

where $\nabla\rho = \left(\frac{\partial\rho}{\partial x}, \frac{\partial\rho}{\partial y}\right)$. Define:

$$R = \frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}), \tag{41}$$

so:

$$\frac{\partial\rho}{\partial t} = R - \frac{\partial(\rho v_x)}{\partial x} - \frac{\partial(\rho v_y)}{\partial y}. \tag{42}$$

Square and integrate:

$$\int\left(\frac{\partial\rho}{\partial t}\right)^2 = \int R^2 + \int(\nabla\cdot(\rho\mathbf{v}))^2 - 2\int\frac{\partial\rho}{\partial t}\nabla\cdot(\rho\mathbf{v}). \tag{43}$$

Apply integration by parts to the cross term:

$$\int\frac{\partial\rho}{\partial t}\nabla\cdot(\rho\mathbf{v})\,dt\,dx\,dy = -\int\rho\frac{\partial}{\partial t}(\nabla\cdot(\rho\mathbf{v}))\,dt\,dx\,dy, \tag{44}$$

since $\int_0^T \frac{\partial}{\partial t}(\rho\nabla\cdot(\rho\mathbf{v}))\,dt = [\rho\nabla\cdot(\rho\mathbf{v})]_0^T = 0$ under periodicity. Expand:

$$\nabla\cdot(\rho\mathbf{v}) = \rho\nabla\cdot\mathbf{v} + \mathbf{v}\cdot\nabla\rho, \tag{45}$$

$$\frac{\partial}{\partial t}(\nabla\cdot(\rho\mathbf{v})) = \frac{\partial\rho}{\partial t}\nabla\cdot\mathbf{v} + \rho\frac{\partial}{\partial t}(\nabla\cdot\mathbf{v}) + \mathbf{v}\cdot\frac{\partial\nabla\rho}{\partial t} + \frac{\partial\mathbf{v}}{\partial t}\cdot\nabla\rho. \tag{46}$$

Assuming $\mathbf{v}$ is smooth and time-independent (short video clips), $\frac{\partial\mathbf{v}}{\partial t} = 0$:

$$\frac{\partial}{\partial t}(\nabla\cdot(\rho\mathbf{v})) = \frac{\partial\rho}{\partial t}\nabla\cdot\mathbf{v} + \mathbf{v}\cdot\frac{\partial\nabla\rho}{\partial t}. \tag{47}$$

Bound:

$$\left|\frac{\partial}{\partial t}(\nabla\cdot(\rho\mathbf{v}))\right| \le |\nabla\cdot\mathbf{v}|\left|\frac{\partial\rho}{\partial t}\right| + V\left|\frac{\partial\nabla\rho}{\partial t}\right|. \tag{48}$$

Thus:

$$\left|\int\rho\frac{\partial}{\partial t}(\nabla\cdot(\rho\mathbf{v}))\right| \le \int|\rho|\left(|\nabla\cdot\mathbf{v}|\left|\frac{\partial\rho}{\partial t}\right| + V\left|\frac{\partial\nabla\rho}{\partial t}\right|\right). \tag{49}$$

Using Cauchy-Schwarz and smoothness ($|\nabla\cdot\mathbf{v}| \le V_D$):

$$\int\left|\frac{\partial\rho}{\partial t}\right|^2 \le \int R^2 + C\int|\nabla\rho|^2, \tag{50}$$

where $C = V^2$ approximates spatial gradient contributions (conservative). Hence:

$$\|\rho\|_{H^1}^2 \le \int\rho^2 + (1+C)\int R^2 + C_1\int|\nabla\rho|^2, \tag{51}$$

with $C_1 = V^2$. Minimizing $\mathcal{L}_{\text{PDE}}$ reduces $\int R^2$, bounding $\|\rho\|_{H^1}^2$ and gradient variance, stabilizing $\hat{V}_{AC}$ (Section III-B). □

Significance: This regularization ensures $\hat{V}_{AC}$'s temporal and spatial consistency (Section III-B), critical for suppressing noise in SpO_2 feature maps under diverse conditions (e.g., lighting changes, Section V-D).

5.5. Joint Optimization of LTC and PDE Loss Enhances Temporal Consistency

Theorem 5. *For the ACuRE model with LTC dynamics $\frac{dx}{dt} = -\left[\frac{1}{\tau} + f(x,I,\theta)\right]x + f(x,I,\theta)A$ and total loss $\mathcal{L}_{total} = \mathcal{L}_{SpO_2} + \alpha\mathcal{L}_{AC\_DC} + \mathcal{L}_{PDE}$, where $\mathcal{L}_{PDE} = \frac{1}{TWH}\int R^2\,dt\,dx\,dy$ and $R = \frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{v})$, the joint optimization reduces the variance of the predicted SpO2 signal gradient:*

$$Var(\nabla Y) \le C\left(\frac{1}{\tau_{eff}} + \|\rho\|_{H^1}^2\right),$$

where $\tau_{eff} = \left(\frac{1}{\tau} + f(x,I,\theta)\right)^{-1}$ is the effective time constant, C is a constant, $\nabla Y = \frac{\partial Y}{\partial t}$, and $Y(t)$ is the predicted SpO2 signal.

Proof. **Setup and Notation:** Let $V(t,x,y) \in \mathbb{R}^{C\times T\times W\times H}$ be the input RGB video, processed through AC/DC convolutional layers to yield $\hat{V}_{AC} = [\rho, v_x, v_y]$, where $\rho(t,x,y) \in \mathbb{R}$ is the blood density field and $\mathbf{v} = (v_x, v_y) \in \mathbb{R}^2$ is the velocity field. The 3D-ResNet-18 backbone extracts features, producing input $I(t) \in \mathbb{R}^D$ to the LTC layer. The LTC dynamics govern the hidden state $x(t) \in \mathbb{R}^D$:

$$\frac{dx}{dt} = -\left[\frac{1}{\tau} + f(x,I,\theta)\right]x + f(x,I,\theta)A, \tag{52}$$

where $f(x, I, \theta) = W \tanh(\gamma I + \mu)$, $\tau \in [0.1, 1.0]$, and $A \in \mathbb{R}^D$. A fully connected layer maps $x(t)$ to $Y(t) = g(x(t))$, with g linear ($W_g \in \mathbb{R}^{1\times D}$). The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SpO}_2} + \alpha \mathcal{L}_{\text{AC\_DC}} + \mathcal{L}_{\text{PDE}}, \tag{53}$$

with $\mathcal{L}_{\text{SpO}_2} = \text{MSE}(Y, Y_{GT}) + \text{NegCorr}(Y, Y_{GT})$, $\mathcal{L}_{\text{AC\_DC}} = \text{MSE}(\hat{V}_{AC}, V_{AC}) + \text{MSE}(\hat{V}_{DC}, V_{DC})$, and $\mathcal{L}_{\text{PDE}} = \frac{1}{TWH} \int R^2 \, dt \, dx \, dy$, $R = \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v})$. **Step 1:**

Gradient of Predicted SpO2 Compute:

$$\begin{aligned} \nabla Y = \frac{dY}{dt} &= W_g \frac{dx}{dt} \\ &= W_g \left(- \left[\frac{1}{\tau} + f(x, I, \theta) \right] x + f(x, I, \theta) A \right). \end{aligned} \tag{54}$$

Define $\tau_{\text{eff}} = \left(\frac{1}{\tau} + f(x, I, \theta) \right)^{-1}$, where $f \in [-|W|, |W|]$, and $\tau \geq 0.1$, so $\tau_{\text{eff}} \in (0, 1]$ for $|W| < 1$.

Step 2: Variance of ∇Y The variance over $t \in [0, T]$ is:

$$\text{Var}(\nabla Y) = \frac{1}{T} \int_0^T (\nabla Y(t) - \bar{Y})^2 \, dt, \quad \bar{Y} = \frac{1}{T} \int_0^T \nabla Y(t) \, dt. \tag{55}$$

Bound:

$$\text{Var}(\nabla Y) \leq \frac{1}{T} \int_0^T (\nabla Y(t))^2 \, dt, \tag{56}$$

where:

$$\nabla Y(t) = W_g \left(- \frac{x(t)}{\tau_{\text{eff}}(t)} + f(x(t), I(t), \theta) A \right). \tag{57}$$

With $\|W_g\|_2 \leq K_g < \infty$:

$$\text{Var}(\nabla Y) \leq K_g^2 \frac{1}{T} \int_0^T \left\| - \frac{x(t)}{\tau_{\text{eff}}(t)} + f(x(t), I(t), \theta) A \right\|^2 dt. \tag{58}$$

Step 3: LTC Contribution From Theorem 1, $x(t)$ is bounded: $\|x(t)\|_2 \leq M_x < \infty$. Also, $f \leq |W|$, $\|A\|_2 \leq M_A$. Split:

$$\begin{aligned} \left\| - \frac{x(t)}{\tau_{\text{eff}}(t)} + f(x(t), I(t), \theta) A \right\|^2 &\leq 2 \left\| \frac{x(t)}{\tau_{\text{eff}}(t)} \right\|^2 \\ &\quad + 2 \| f(x(t), I(t), \theta) A \|^2, \end{aligned} \tag{59}$$

so:

$$\text{Var}(\nabla Y) \leq K_g^2 \frac{1}{T} \int_0^T \left(2 \frac{M_x^2}{\tau_{\text{eff}}^2(t)} + 2 |W|^2 M_A^2 \right) dt. \tag{60}$$

Let $C_1 = 2K_g^2 M_x^2$, $C_2 = 2K_g^2 |W|^2 M_A^2$:

$$\text{Var}(\nabla Y) \leq C_1 \mathbb{E} \left[\frac{1}{\tau_{\text{eff}}^2} \right] + C_2. \tag{61}$$

Step 4: PDE Contribution $\mathcal{L}_{\text{PDE}}$ bounds:

$$\|\rho\|_{H^1}^2 = \int \left(\rho^2 + \left| \frac{\partial \rho}{\partial t} \right|^2 + |\nabla \rho|^2 \right) dt \, dx \, dy \leq C_3 + C_4 \int R^2, \tag{62}$$

reducing $\|\rho\|_{H^1}^2$ as $\mathcal{L}_{\text{PDE}}$ decreases. For $I(t) = h(\hat{V}_{AC}(t))$, Lipschitz h (constant L_h) gives:

$$\|I(t) - I(s)\|_2 \leq L_h C_6 \|\rho\|_{H^1} \sqrt{|t - s|}, \tag{63}$$

smoothing $I(t)$ and thus $\frac{dx}{dt}$.

Step 5: Joint Bound Combine:

$$\text{Var}(\nabla Y) \leq C_1 \mathbb{E} \left[\frac{1}{\tau_{\text{eff}}^2} \right] + C_2 + C_7 \|\rho\|_{H^1}^2, \tag{64}$$

approximating $\mathbb{E} \left[\frac{1}{\tau_{\text{eff}}^2} \right] \approx \left(\mathbb{E} \left[\frac{1}{\tau_{\text{eff}}} \right] \right)^2$. Define $C = \max(C_1, C_7)$:

$$\text{Var}(\nabla Y) \leq C \left(\frac{1}{\tau_{\text{eff}}} + \|\rho\|_{H^1}^2 \right) + C_8. \tag{65}$$

As $\mathcal{L}_{\text{total}} \to 0$, the bound tightens, proving the theorem. □

6. Dataset Details

We conducted our SpO_2 estimation experiments on three publicly available datasets: PURE [19], BH-rPPG [22], and VIPL-HR [15]. The PURE[19] dataset comprises 60 face videos from 10 subjects recorded during six one-minute sessions featuring various head motions (steady, talking, and translations).The videos were captured at a resolution of 640×480 pixels with a frame rate of 30 FPS, while SpO_2 signals were recorded using a finger pulse oximeter operating at 60 Hz. The BH-rPPG[22] dataset contains 105 videos from 35 participants, collected under three distinct lighting conditions: low (8 lux), medium (42.4 lux), and high (104 lux). These videos were recorded at 640×480 resolution with a frame rate of 15 FPS, with SpO_2 measurements obtained from a CONTEC CMS50E pulse oximeter. The VIPL-HR[15] dataset comprises 2,378 visible light (VIS) and 752 near-infrared (NIR) videos from 107 subjects, captured under varying head motions, lighting conditions, and multiple acquisition devices, including Logitech C310, RealSense F200, and HUAWEI P9. The ground-truth physiological signals were recorded using a CONTEC CMS60C BVP sensor. These datasets collectively provide a diverse benchmark for assessing the accuracy and generalizability of SpO_2 estimation models.

7. Error Metric Formula

- **Mean Absolute Error (MAE):**

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |Y_i - \hat{Y}_i|, \tag{66}$$

where Y_i and $\hat{Y}_i$ represent the ground truth and predicted SpO_2 values, respectively, and N is the total number of samples.

- **Root Mean Squared Error (RMSE):**

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}. \tag{67}$$

RMSE provides greater weight to larger errors, making it more sensitive to significant deviations.

- **Pearson Correlation Coefficient (r):**

$$r = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \sum_{i=1}^{N}(\hat{Y}_i - \bar{\hat{Y}})^2}}, \tag{68}$$

where $\bar{Y}$ and $\bar{\hat{Y}}$ are the mean values of the ground truth and predicted SpO_2, respectively. A value of r closer to 1 indicates stronger agreement between predictions and ground truth.

References

[1] Yusuke Akamatsu, Yoshifumi Onishi, and Hitoshi Imaoka. Blood oxygen saturation estimation from facial video via dc and ac components of spatio-temporal map. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[2] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 1

[3] Joao Carreira et al. Video vision transformers: A survey. *arXiv preprint arXiv:2201.04891*, 2022. Adapted for 3D ViT variants in video tasks. 2, 3

[4] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31:6571–6583, 2018. 1

[5] Chun-Hong Cheng, Zhikun Yuen, Shutao Chen, Kwan-Long Wong, Jing-Wei Chin, Tsz-Tai Chan, and Richard H. Y. So. Contactless blood oxygen saturation estimation from facial videos using deep learning. *Bioengineering*, 11(3):251, 2024. 5

[6] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 1

[7] 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. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[8] Ken-ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural networks*, 6(6):801–806, 1993. 1

[9] Ramin Hasan, Christian Toth, Wei Hu, Kartik Srinivasan, Chung Tran, and David A. Clifton. Liquid time-constant networks. *arXiv preprint arXiv:2006.04439*, 2021. 1

[10] R. Hasani, M. Lechner, A. Amini, D. Rus, and R. Grosu. Liquid time-constant networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7657–7666, 2021. 4, 5, 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 3

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1

[13] Min Hu, Xia Wu, Xiaohua Wang, Yan Xing, Ning An, and Piao Shi. Contactless blood oxygen estimation from face videos: A multi-model fusion method based on deep learning. *Biomedical Signal Processing and Control*, 81:104487, 2023. 5

[14] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face videos. *arXiv preprint arXiv:1810.04927*, 2018. 1

[15] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 562–576. Springer, 2019. 8

[16] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Heart rate measurement based on 3d central difference convolutional network for remote photoplethysmography. *IEEE Transactions on Biomedical Engineering*, 68(6):1925–1936, 2021. 2

[17] Jiahe Peng, Weihua Su, Haiyong Chen, Jingsheng Sun, and Zandong Tian. Cl-spo2net: Contrastive learning spatiotemporal attention network for non-contact video-based spo_2 estimation. *Bioengineering*, 11(2):113, 2024. 5

[18] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. 1

[19] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014. 1, 8

[20] Xiantao Sun, Tao Wen, Weihai Chen, and Bin Huang. Cc-spo2net: Camera-based contactless oxygen saturation measurement foundation model in clinical settings. *IEEE Transactions on Instrumentation and Measurement*, 73:4005211, 2024. 5

[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[22] Ze Yang, Haofei Wang, and Feng Lu. Assessment of deep learning-based heart rate estimation using remote photoplethysmography under different illuminations. *IEEE Transactions on Human-Machine Systems*, 52(6):1236–1246, 2022. 8

[23] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 2, 3

[24] Yujia Zhang, Zijun Wang, Jian Yang, and Xinyu Wang. Bh-rppg: A benchmark dataset for heart rate estimation from face videos. *arXiv preprint arXiv:1906.09234*, 2019. 1