FUSE: Filter-Free Unified Spatiotemporal Estimation of SpO_2 via Wave-Transport Modeling

Shahzad Ahmad[1], Nuvvuru Sai Charan Reddy[2], Ram Prasad Padhy[2], Sukalpa Chanda[1], and Umapada Pal[3]

[1] Østfold University of Applied Sciences, Norway
[2] Indian Institute of Technology Bhubaneswar, India
[3] Indian Statistical Institute, Kolkata
shahzaa@hiof.no

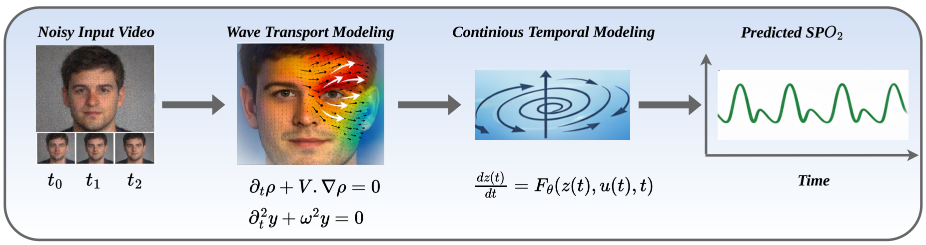

Fig. 1: Overview of the proposed framework for filter-free SpO_2 estimation from facial videos. Pulsatile dynamics are modeled as a coupled wave-transport system using harmonic ODE constraints and transport-based PDE regularization, while continuous-time neural dynamics estimate the final SpO_2 signal.

Abstract. Blood oxygen saturation (SpO_2) is a vital indicator of respiratory and cardiovascular health, increasingly targeted by camera-based, contactless monitoring systems. Existing video-based approaches typically rely on predefined frequency filtering to separate pulsatile (AC) and baseline (DC) components. Such spectral decomposition propagates band-limited noise and assumes signal stationarity, limiting robustness under motion, illumination variation, and subject diversity.

We propose **FUSE**, a filter-free, physics-guided framework that learns AC and DC representations via structured differential constraints rather than static spectral filtering. The pulsatile component is regularized as a band-limited harmonic field through a second-order ordinary differential equation, while spatial coherence is enforced by modeling it as a transported density field under an advection-based partial differential constraint. This coupled wave-transport formulation enforces physiologically consistent spatiotemporal dynamics without explicit frequency-domain operations.

Integrated into a lightweight spatiotemporal backbone with continuous-time temporal modeling, FUSE achieves improved robustness and consistently lower estimation errors across multiple public benchmarks, establishing a principled and physics-driven alternative for contactless SpO_2 estimation from facial videos. **Code is available:** link

Keywords: SpO_2 Estimation · Remote Photoplethysmography · Physics-Informed Learning · Wave-Transport Modeling · Continuous-Time Neural Networks

1 Introduction

Blood oxygen saturation (SpO_2) is a critical vital sign for assessing respiratory and circulatory health and enables early detection of hypoxemia in conditions such as chronic obstructive pulmonary disease (COPD), heart failure, and acute respiratory distress syndrome (ARDS). With over 300 million COPD cases globally [29] and the telemedicine market projected to exceed $450 billion by 2026 [19], scalable and non-contact SpO_2 monitoring has become increasingly important. Traditional contact-based pulse oximeters [5] can cause discomfort, increase infection risk [6], and are impractical for large-scale continuous monitoring. Camera-based remote photoplethysmography (rPPG) offers a hygienic and wearable-free alternative, with growing relevance for telemedicine and remote healthcare applications [11, 24].

The classical principle of SpO_2 estimation relies on separating pulsatile (AC) and baseline (DC) components of reflected light. The AC component captures blood-volume oscillations driven by cardiac activity, while the DC component reflects static tissue absorption [28]. Conventional ratio-of-ratios (RoR) formulations [16, 27] require dedicated infrared wavelengths, which limits compatibility with standard RGB cameras. In camera-based approaches, AC/DC separation is typically approximated using frequency-domain filtering [22]. However, bandpass filtering propagates noise within the selected passband, assumes stationarity, and degrades under motion or illumination changes. Moreover, discrete recurrent models [14] are often used for temporal modeling, although physiological signals evolve in continuous time, leading to instability under varying frame rates or dynamic perturbations [4].

To address these limitations, we propose **FUSE** (**F**ilter-free **U**nified **S**patiotemporal **E**stimation), a physics-guided framework that models SpO_2 estimation as a coupled wave-transport dynamical system (Fig. 1). Instead of relying on predefined frequency filtering, FUSE learns pulsatile and baseline representations directly through differential constraints. The pulsatile component is modeled as a band-limited harmonic field governed by a second-order ordinary differential equation (ODE), encouraging physiologically plausible oscillatory dynamics. To capture spatial coherence across the facial surface, the AC representation is further interpreted as a transported scalar density and regularized via an advection-based partial differential equation (PDE) under a near-incompressible assumption. Together, these constraints enforce both local oscillatory structure (wave behavior) and coherent spatial propagation (transport behavior) in the learned representation.

To accommodate continuous physiological dynamics, FUSE incorporates continuous time temporal models, including Liquid Time-Constant (LTC) networks [13], Continuous-Time RNNs (CTRNN) [12], and Neural Ordinary Differential Equations (Neural-ODE) [7]. Unlike fixed-step recurrent architectures, continuous-time models naturally adapt to varying frame rates across datasets (PURE, VIPL-HR, BH-rPPG) and improve cross-dataset generalization. Importantly, the proposed differential formulation is backbone-agnostic, remaining effective across both convolutional and transformer architectures.

Extensive experiments on PURE, BH-rPPG, and VIPL-HR demonstrate that FUSE achieves state-of-the-art performance in both intra-dataset and cross-dataset settings, with improved robustness to motion and illumination variations. These results establish filter-free wave-transport modeling as a principled approach for camera-based physiological sensing.

Our contributions are summarized as follows:

- We introduce a ***filter-free differential formulation for AC/DC learning***, replacing frequency-domain decomposition with harmonic ODE constraints and transport-based PDE regularization.
- We propose a ***wave transport framework*** that jointly models oscillatory temporal dynamics and spatial propagation of pulsatile signals, improving robustness to motion and illumination variations.
- We integrate the formulation with ***continuous-time neural dynamics*** (LTC, CTRNN, Neural-ODE), enabling stable temporal modeling and improved cross-dataset generalization.

FUSE unifies physiological modeling, differential constraints, and continuous-time learning into a framework for reliable non-contact SpO_2 estimation in next-generation telemedicine systems.

2 Related Work

Remote SpO_2 estimation from facial videos has progressed from signal-processing-based ratio formulations to deep spatiotemporal learning frameworks. Despite recent advances, robust AC/DC separation, temporal stability, and cross-dataset generalization remain open challenges.

Contactless and Learning-Based SpO_2 Estimation: Early contactless approaches adapted the classical ratio-of-ratios principle to RGB cameras through signal normalization and temporal filtering [16]. These camera-based adaptations estimate the AC/DC components by using the temporal mean as the DC component and either temporal standard deviation or band-pass filtering to isolate the AC component [22, 23].CNN-based systems enable smartphone-based SpO_2 estimation [18], while 3D CNNs with contrastive objectives improve feature representation [21]. Recent deep learning methods directly learn nonlinear spatiotemporal mappings from video to SpO_2. Sun et al. [26] proposed $CCSpO_2$, leveraging convolutional modeling for camera-based estimation. Hu et al. [15] introduced a contactless framework with enhanced feature learning for improved robustness. ROSE-Net [10] and SHINE [1] further improved representation learning through refined spatiotemporal encoders. ITSCAN [33] and Cheng et al. [8] focused on improving temporal aggregation and generalization across varying acquisition conditions. While effective, most of these methods rely on implicit frequency filtering or discrete temporal modeling [9, 14], which assumes fixed sampling rates and may degrade under motion or illumination variation.

Physics-Aware and Structured Modeling: To improve physiological consistency, recent works incorporate structured priors into learning. Akamatsu et al. [4] modeled blood volume dynamics to enhance waveform fidelity. PHASE [3] introduced phase-aware modeling to better preserve pulsatile structure during estimation. ACuRE [2] combined dual-branch AC/DC separation with Liquid Time-Constant (LTC) temporal modeling and a continuity-inspired PDE regularizer to enforce waveform consistency. Although ACuRE demonstrated improved robustness, its AC/DC supervision relied on filtered pseudo-labels and waveform-level constraints.

Continuous-Time Neural Dynamics: Continuous-time neural models provide a principled framework for modeling physiological dynamics. Liquid Time-Constant (LTC) networks [13] introduce neuron-wise ODEs with adaptive time constants, enabling stable modeling of fine-grained temporal behavior. Continuous-Time Recurrent Neural Networks (CTRNN) [12] provide universal approximation guarantees for dynamical systems, while Neural Ordinary Differential Equations (Neural-ODE) [7] formulate hidden state evolution as solutions to learned differential equations. These approaches naturally handle irregular sampling and varying frame rates, making them well-suited for rPPG-based physiological estimation.

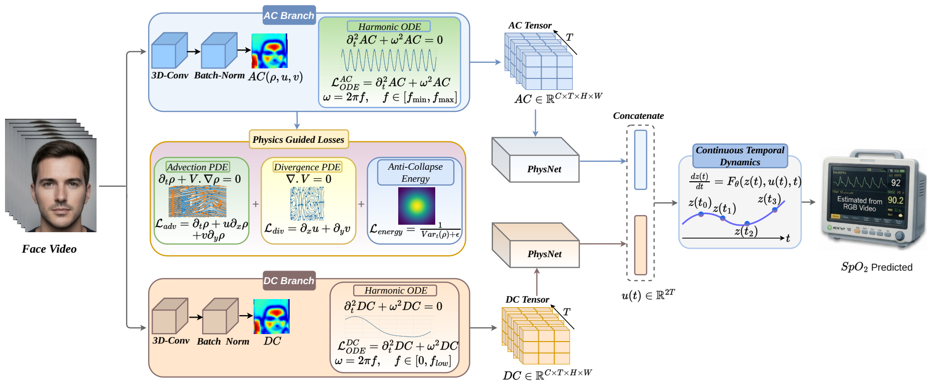

Fig. 2: Overview of the proposed FUSE framework for camera-based SpO_2 estimation. The method learns pulsatile (AC) and baseline (DC) representations from facial videos without explicit frequency filtering. Temporal oscillatory dynamics are enforced through harmonic ODE constraints, while spatial coherence of pulsatile signals is modeled via a transport-based PDE. The resulting representations are fused and processed by a continuous-time neural dynamical system to produce frame-wise SpO_2 predictions.

3 Proposed Method

We propose a filter-free, physics-informed framework for supervised SpO_2 estimation from facial RGB videos. In contrast to ACuRE [2], which relied on frequency-domain filtering to generate AC/DC pseudo-ground-truth signals, the proposed approach eliminates static spectral decomposition and instead enforces physiologically grounded differential constraints within an end-to-end continuous-time modeling framework. Given an input RGB facial video $\mathbf{V} \in \mathbb{R}^{B\times 3\times T\times H\times W}$, where $W \times H$ are frame dimensions and T is the frame count, our goal is to estimate the frame-wise blood oxygen saturation waveform $\hat{\mathbf{y}} \in \mathbb{R}^{B\times T}$ The proposed architecture integrates three tightly coupled components:

- Physics-guided spatio-temporal AC/DC representation learning via differential equation regularization (harmonic ODE + transport PDE)
- Generalized continuous-time temporal dynamics modeling
- Direct end-to-end SpO_2 waveform supervision

We first revisit the physiological foundation underlying SpO_2 estimation.

3.1 Preliminary: Ratio-of-Ratios

SpO_2 estimation is classically derived from the ratio-of-ratios (RoR) principle, which exploits the differential optical absorption of oxygenated (HbO_2) and deoxygenated (Hb) hemoglobin at distinct wavelengths [22]. Under a small-variation Beer-Lambert approximation, the reflected intensity at wavelength λ is $I_\lambda(t) = I_{0,\lambda} \exp\big(-\epsilon_\lambda c(t) d\big)$, where ϵ_λ is the extinction coefficient, $c(t)$ is time-varying hemoglobin concentration, and d is the optical path length.

Over short temporal windows, the signal decomposes as $I_\lambda(t) = DC_\lambda + AC_\lambda(t)$, with $DC_\lambda = \mathbb{E}_t[I_\lambda(t)]$ and $AC_\lambda(t) = I_\lambda(t) - DC_\lambda$. The ratio-of-ratios is then defined as $\mathrm{RoR} = \frac{AC_{\lambda_1}/DC_{\lambda_1}}{AC_{\lambda_2}/DC_{\lambda_2}}$, and SpO_2 is approximately affine, $\mathrm{SpO}_2 \approx A.\,\mathrm{RoR} + B$.

In RGB camera settings, true infrared bands are unavailable and RGB channels serve as proxy spectra. Conventional non-contact methods estimate AC/DC components via frequency filtering [22], implicitly assuming temporal stationarity and spectral separability between pulse and noise. Instead of relying on filtered pseudo-labels, we learn AC/DC representations directly by imposing differential constraints that encode their intrinsic oscillatory and transport dynamics. A detailed derivation of the RoR formulation under linearized Beer-Lambert assumptions is provided in the ***Supplementary Sec: 7*** .

3.2 Overview of the Proposed Architecture

The proposed framework models physiological dynamics at three hierarchical levels: (i) temporal oscillation of pulsatile signals, (ii) spatial transport of pulsatile density across the facial surface, and (iii) global latent evolution of the physiological waveform. An overview of the architecture is illustrated in Fig. 2.

Given an input facial video

$$\mathbf{V} \in \mathbb{R}^{B\times 3\times T\times H\times W}$$

the model first extracts two complementary spatio–temporal feature fields, representing pulsatile (AC) and baseline (DC) components. Unlike conventional approaches that rely on frequency filtering, these components are learned directly through differential constraints.

Temporal Oscillation Modeling (ODE Level): At the lowest level, we enforce physiologically plausible temporal oscillations on the AC and DC representations through harmonic differential constraints. Two parallel spatio-temporal blocks process the input video and produce feature tensors

$$\hat{\mathbf{X}}_{AC}, \hat{\mathbf{X}}_{DC} \in \mathbb{R}^{B \times C \times T \times H \times W}$$

Both branches operate with unit spatial and temporal stride, thus preserving the full spatio-temporal resolution of the input.

Instead of supervising these tensors with bandpass/lowpass filtered pseudo-targets, their temporal dynamics are regularized using harmonic ordinary differential equations. Specifically, AC features are constrained to follow band-limited oscillatory dynamics, while DC features are encouraged to remain slowly varying. This formulation removes reliance on frequency-domain decomposition and encourages physiologically consistent temporal structure (Section 3.3).

Spatial Transport Modeling (PDE Level): At the intermediate level, we model spatial propagation of pulsatile signals across the facial surface. The AC feature field is interpreted as a transported scalar density $\rho(t, x, y)$ with an associated latent velocity field $\mathbf{v}(t, x, y)$.

Under this interpretation, pulsatile propagation is governed by a transport equation derived from mass conservation principles:

$$\frac{\partial \rho}{\partial t} + \mathbf{v} \cdot \nabla \rho \approx 0 \tag{1}$$

This advection constraint enforces spatial coherence of pulsatile flow, while a divergence regularization term stabilizes the learned velocity field by encouraging near-incompressible transport. The full PDE formulation and loss functions are described in Section 3.4.

Feature Extraction and Representation Fusion: After differential regularization, the AC and DC feature tensors are processed by two parallel PhysNet backbones [31], a 3D convolutional architecture designed for physiological signal extraction. Each backbone maps the spatio-temporal tensors into compact temporal embeddings

$$\mathbf{f}_{AC}(t), \mathbf{f}_{DC}(t) \in \mathbb{R}^{D}$$

which capture pulsatile and baseline dynamics respectively. These embeddings are concatenated to form the fused representation

$$\mathbf{u}(t) = [\mathbf{f}_{AC}(t), \mathbf{f}_{DC}(t)] \in \mathbb{R}^{2D} \tag{2}$$

Global Continuous-Time Dynamics: At the highest level, the fused features drive a latent continuous-time dynamical system that models the evolution of the physiological waveform. Let $\mathbf{z}(t) \in \mathbb{R}^d$ denote the latent physiological state. Its temporal evolution is governed by

$$\frac{d\mathbf{z}(t)}{dt} = F_\theta\big(\mathbf{z}(t), \mathbf{u}(t), t\big) \tag{3}$$

where $F_\theta(\cdot)$ is a learnable neural vector field. The predicted SpO_2 signal is obtained through

$$\hat{y}(t) = W_o \mathbf{z}(t) + b_o \tag{4}$$

Equation (3) provides a unified continuous-time formulation that encompasses several temporal modeling paradigms, including Neural ODEs [7], continuous-time recurrent neural networks (CTRNN) [12], and Liquid Time-Constant networks (LTC) [13]. Unlike discrete recurrence, this formulation naturally handles variable frame rates and irregular temporal sampling.

Together, harmonic ODE constraints, transport-based PDE regularization, and continuous-time latent dynamics form a hierarchical physics-guided framework for robust SpO_2 estimation. Detailed mathematical derivations are provided in the ***Supplementary Sec:10***.

3.3 Differential-Equation-Based AC/DC Formulation

Prior work (e.g., ACuRE [2]) supervises AC/DC decomposition using bandpass/lowpass filtered pseudo-targets, which inevitably preserve in-band noise. In contrast, we learn AC and DC representations without filtered targets by imposing time-domain dynamical constraints directly on the learned feature fields.

Let $\hat{\mathbf{X}}_{AC}, \hat{\mathbf{X}}_{DC} \in \mathbb{R}^{B\times C\times T\times H\times W}$ denote the AC and DC feature tensors produced by the two branches (Section 3.2). For each batch index b, channel c, and spatial coordinate (x, y), we view the temporal traces $\hat{x}^{(c)}_{AC}(t) \doteq \hat{x}^{(c)}_{AC}(t, x, y)$ and $\hat{x}^{(c)}_{DC}(t) \doteq \hat{x}^{(c)}_{DC}(t, x, y)$ as continuous-time signals sampled at frame times $t \in \{1, \dots, T\}$.

Band-Limited Harmonic Constraint for AC: A sinusoid satisfies the harmonic oscillator equation $\frac{d^2 s}{dt^2} + \omega^2 s = 0$. Motivated by the quasi-periodic nature of pulsatile dynamics, we constrain AC features to behave as *band-limited oscillators*. Let $\{\omega_k\}_{k=1}^K$ be a set of angular frequencies sampled from a physiologically plausible heart-rate band $\omega_k \in [\omega_{\min}, \omega_{\max}]$ (with $\omega_k = 2\pi f_k$). We define the AC harmonic loss as

$$\mathcal{L}^{AC}_{ODE} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{b,c,t,x,y}\left[\phi\Big(\partial_t^2 \hat{x}^{(c)}_{AC}(t, x, y) + \omega_k^2 \hat{x}^{(c)}_{AC}(t, x, y)\Big)\right] \tag{5}$$

where $\phi(\cdot)$ is a robust penalty (e.g., Charbonnier or ℓ_1).

Low-Frequency Constraint for DC: The DC component corresponds to slowly varying baseline absorption. We enforce this by applying the same harmonic prior but sampling frequencies from a low-frequency interval $\omega_k \in [0, \omega_{\text{low}}]$, yielding

$$\mathcal{L}_{ODE}^{DC} = \frac{1}{K_{dc}} \sum_{k=1}^{K_{dc}} \mathbb{E}_{b,c,t,x,y} \left[\phi\Big(\partial_t^2 \hat{x}_{DC}^{(c)}(t,x,y) + \omega_k^2 \hat{x}_{DC}^{(c)}(t,x,y) \Big) \right] \tag{6}$$

Why ODE Supervision Instead of Filtering? If the observed temporal signal is $V(t) = S(t) + N(t)$, frequency filtering yields $V_{AC}(t) = S_{AC}(t) + N_{AC}(t)$, thereby preserving noise inside the passband. In contrast, Eqs. (5)-(6) regularize the *intrinsic dynamics* of the learned representation: AC features are encouraged to follow band-limited oscillatory dynamics, while DC features remain smooth and low-frequency. This provides a filter-free alternative to frequency-domain decomposition. A detailed theoretical connection between harmonic residual minimization and soft bandpass behavior is provided in the ***Supplementary Sec:8***.

3.4 PDE Constraint: Transported Pulsatile Density

While the harmonic ODE constraint regularizes temporal oscillations, pulsatile signals also exhibit spatial propagation across the facial surface due to blood flow and optical diffusion. To capture this behavior, we model the AC feature field as a transported scalar density governed by a transport partial differential equation.

Let the AC tensor be

$$\hat{\mathbf{X}}_{AC} \in \mathbb{R}^{B \times 3 \times T \times H \times W}$$

We interpret its channels as

$$\rho(t,x,y) = \hat{\mathbf{X}}_{AC}^{(1)}(t,x,y) \tag{7}$$

$$\mathbf{v}(t,x,y) = (u(t,x,y),\, v(t,x,y)) = \left(\hat{\mathbf{X}}_{AC}^{(2)}, \hat{\mathbf{X}}_{AC}^{(3)} \right) \tag{8}$$

where ρ represents the transported pulsatile density and $\mathbf{v}$ denotes a latent velocity field that models spatial propagation of pulsatile signals.

Mass Conservation: Starting from conservation of transported density, the continuity equation is

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{9}$$

In facial skin tissue, pulsatile propagation is dominated by transport rather than compression. We therefore assume near-incompressible flow

$$\nabla \cdot \mathbf{v} \approx 0 \tag{10}$$

which reduces the continuity equation to the advection equation

$$\frac{\partial \rho}{\partial t} + \mathbf{v} \cdot \nabla \rho = 0 \tag{11}$$

Advection Residual: To enforce this constraint, we compute the advection residual

$$R_{adv}(t, x, y) = \frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + v\frac{\partial \rho}{\partial y} \tag{12}$$

The transport regularization loss is defined as

$$\mathcal{L}_{adv} = \mathbb{E}_{t,x,y}\left[\phi(R_{adv})\right] \tag{13}$$

where $\phi(\cdot)$ denotes a robust penalty function.

Divergence Regularization: To stabilize the learned velocity field, we softly enforce the incompressibility condition using

$$\mathcal{L}_{div} = \mathbb{E}_{t,x,y}\left[\phi(\nabla \cdot \mathbf{v})\right] \tag{14}$$

Anti-Collapse Constraint: Equation (11) admits trivial constant solutions ($\rho = \text{const}$). To avoid degenerate solutions, we enforce a minimum temporal energy in the density field:

$$\mathcal{L}_{energy} = \mathbb{E}_{x,y}\left[\frac{1}{\mathrm{Var}_t(\rho) + \epsilon}\right] \tag{15}$$

This constraint prevents density collapse and encourages non-trivial pulsatile dynamics.

Together, the proposed PDE regularization enforces spatio-temporal coherence of pulsatile transport. Combined with the harmonic ODE supervision, it enables learning physiologically meaningful AC representations without relying on frequency-domain filtering. A detailed derivation of the transport formulation and its connection to physiological signal propagation is provided in the ***Supplementary Sec:9***.

3.5 $\mathbf{SpO_2}$ Supervision Loss

Given the ground-truth waveform $\mathbf{y} \in \mathbb{R}^{B \times T}$ and the predicted signal $\hat{\mathbf{y}}$, we supervise the model using a regression objective that combines amplitude fidelity and waveform similarity:

$$\mathcal{L}_{SpO_2} = \mathrm{MSE}(\hat{\mathbf{y}}, \mathbf{y}) - \lambda_{corr}\,\mathrm{Corr}(\hat{\mathbf{y}}, \mathbf{y}) \tag{16}$$

where $\mathrm{Corr}(\cdot, \cdot)$ denotes differentiable Pearson correlation and λ_{corr} controls the trade-off between signal magnitude accuracy and temporal waveform alignment.

3.6 Overall Training Objective

The final objective combines waveform supervision with the proposed differential regularization terms:

$$\mathcal{L}_{total} = \mathcal{L}_{SpO_2} + \alpha \left(\mathcal{L}_{ODE}^{AC} + \mathcal{L}_{ODE}^{DC} \right) + \lambda_{adv} \mathcal{L}_{adv} + \lambda_{div} \mathcal{L}_{div} + \lambda_{energy} \mathcal{L}_{energy} \quad (17)$$

Here α, λ_{adv}, λ_{div}, and λ_{energy} control the strength of harmonic temporal regularization, spatial transport consistency, velocity incompressibility, and anti-collapse constraints respectively.

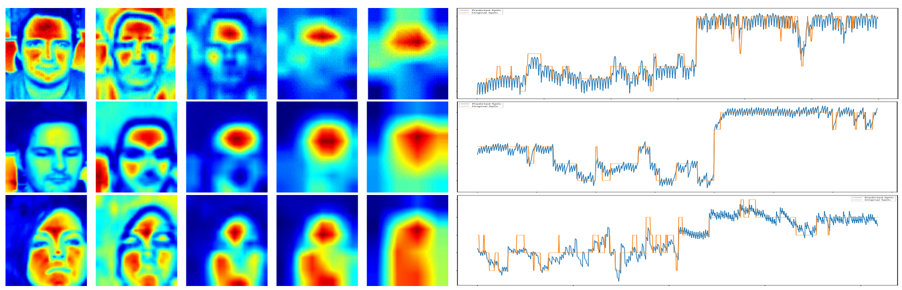

Fig. 3: Visualization of intermediate AC-branch feature maps and corresponding SpO_2 predictions on the PURE dataset [25]. Feature maps from successive PhysNet layers (32×32 to 2×2) illustrate multi-scale pulsatile representation learning. The right plots show predicted (blue) and ground-truth (orange) SpO_2 signals over time, demonstrating strong temporal alignment.

4 Training Details

Datasets: We evaluate FUSE on three publicly available SpO_2 benchmarks: PURE [25], BH-rPPG [30], and VIPL-HR [20]. These datasets differ in frame rate, illumination, subject diversity, and acquisition devices, providing a comprehensive evaluation of cross-condition robustness. Detailed dataset statistics are provided in ***Supplementary Sec:11***.

Preprocessing: Facial landmarks are extracted using OpenFace [32]. The detected 68 landmarks are used to align and crop facial regions to reduce motion-induced variability. Aligned face patches are resized to 32×32 and segmented into clips of length $T = 120$ frames. This preprocessing ensures spatial consistency while preserving pulsatile dynamics.

Implementation Details: FUSE is implemented in PyTorch and trained for 60 epochs using AdamW [17] with learning rate 10^{-4} and batch size 20. We employ 5-fold subject-disjoint cross-validation (3 folds training, 1 validation, 1 testing). Results are reported as mean $\pm$ standard deviation. Continuous-time temporal modeling (LTC, CTRNN, Neural-ODE) is trained jointly with harmonic AC/DC ODE constraints and transport-based PDE regularization.

Model Complexity: Under a fixed PhysNet backbone, forward computation is dominated by the backbone (∼50 GFLOPs/clip), while temporal modules (LTC/CTRNN/Neural-ODE) add minimal parameters and negligible FLOPs. The proposed PDE regularization is parameter-free and applied only during training, introducing no inference-time overhead. Training cost increases due to derivative computation on spatio-temporal feature maps. A detailed breakdown of parameters, FLOPs, latency, and cross-paper comparisons is provided in ***Supplementary Sec:5***.

Evaluation Metrics: We report mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient (r). All metrics are computed over 120-frame windows at their respective dataset frame rates (30 fps, 25 fps, 15 fps). Formal definitions are provided in ***Supplementary Sec:S3***. Lower MAE/RMSE and higher r indicate better performance.Detailed are provided in ***Supplementary Sec:12***.

5 Results and Ablation

5.1 Intra-Dataset Evaluation and Qualitative Analysis

Table 1 reports SpO_2 estimation performance on the PURE, BH-rPPG, and VIPL-HR benchmarks. We compare FUSE with recent camera-based SpO_2 estimation methods, including PHASE [3], ACuRE [2], and other state-of-the-art approaches [1, 8, 10, 15, 26, 33]. FUSE consistently outperforms prior methods across all datasets. On PURE, the MAE is reduced from 0.28 (ACuRE) to 0.15. On BH-rPPG, the CTRNN variant achieves the best performance (MAE 0.24, $r = 0.94$). On VIPL-HR, FUSE maintains strong robustness under more challenging conditions, reducing MAE from 0.72 to 0.56. These results demonstrate that the proposed differential formulation, combining harmonic ODE modeling with transport-based PDE constraints, improves both estimation accuracy and cross-dataset robustness over frequency-filter-based decomposition strategies. To further analyze the learned representations, Fig. 3 visualizes intermediate AC-branch feature maps and corresponding SpO_2 predictions. Feature maps from successive PhysNet layers progressively highlight pulsatile structures, while the predicted signals closely follow the ground-truth waveform, supporting the effectiveness of the proposed formulation.

5.2 Robustness Under Motion and Illumination Variations

We evaluate robustness under controlled perturbations using condition-wise analysis on PURE and illumination-specific analysis on BH-rPPG.

Table 2 reports MAE under different scenarios on PURE. FUSE maintains low estimation error across all conditions. The LTC variant achieves the best performance under steady conditions (MAE 0.12), while Neural-ODE shows strong robustness under head movement (0.11). Overall, the limited performance degradation under motion and lighting changes indicates stable temporal dynamics and coherent spatial modeling.

Table 1: Intra-dataset SpO_2 estimation results on PURE, BH-rPPG, and VIPL-HR. MAE and RMSE are reported in percentage units, and r denotes Pearson correlation. $\uparrow$ indicates higher is better and $\downarrow$ indicates lower is better.

Method	PURE			BH-rPPG			VIPL-HR		
	MAE $\downarrow$	RMSE $\downarrow$	$r \uparrow$	MAE $\downarrow$	RMSE $\downarrow$	$r \uparrow$	MAE $\downarrow$	RMSE $\downarrow$	$r \uparrow$
Sun et al. [26]	0.54	0.68	–	–	–	–	–	–	–
Hu et al. [15]	0.63	0.80	–	–	–	–	1.00	1.43	–
ROSE-Net [10]	1.95	2.46	–	–	–	–	–	–	–
SHINE [1]	1.13	1.46	–	–	–	–	–	–	–
ITSCAN [33]	–	–	–	–	–	–	1.10	1.19	–
Cheng et al. [8]	–	–	–	–	–	–	0.97	1.20	–
PHASE [3]	0.53 ± 0.10	0.72 ± 0.12	0.89 ± 0.03	0.97 ± 0.04	0.91 ± 0.35	0.82 ± 0.14	–	–	–
Akamatsu et al. [4]	0.61 ± 0.16	0.75 ± 0.13	0.88 ± 0.04	1.90 ± 0.56	1.49 ± 0.21	0.42 ± 0.22	–	–	–
ACuRE [2]	0.28 ± 0.17	0.50 ± 0.16	0.95 ± 0.03	0.39 ± 0.18	0.61 ± 0.13	0.92 ± 0.04	0.72 ± 0.20	0.84 ± 0.12	0.87 ± 0.04
FUSE (Neural-ODE)	0.17 ± 0.07	0.40 ± 0.08	0.89 ± 0.09	0.30 ± 0.12	0.54 ± 0.10	0.93 ± 0.02	0.60 ± 0.27	0.76 ± 0.16	0.90 ± 0.05
FUSE (CTRNN)	**0.15 ± 0.05**	**0.38 ± 0.07**	0.90 ± 0.08	**0.24 ± 0.05**	**0.49 ± 0.05**	**0.94 ± 0.01**	**0.56 ± 0.17**	**0.74 ± 0.12**	0.91 ± 0.03
FUSE (LTC)	**0.15 ± 0.07**	**0.38 ± 0.08**	0.92 ± 0.05	0.28 ± 0.07	0.53 ± 0.07	0.93 ± 0.03	0.57 ± 0.11	0.75 ± 0.08	0.91 ± 0.03

Table 3 evaluates robustness under varying illumination levels on BH-rPPG. Across low-, medium-, and high-lux regimes, FUSE exhibits stable performance. The CTRNN variant achieves the lowest MAE across all lighting levels, including challenging low-light conditions (8 lux).

These results indicate that the proposed wave-transport formulation, combining harmonic ODE constraints with transport-based PDE regularization, improves robustness to motion artifacts and illumination variability in camera-based physiological sensing.

Table 2: Condition-wise MAE ($\downarrow$) of FUSE variants on the PURE dataset under steady, motion, and illumination perturbations (mean ± std).

Method	Steady	Talking/Laughing	Body Movement	Dark Light	Head Movement
FUSE (LTC)	**0.12 ± 0.06**	**0.16 ± 0.03**	**0.20 ± 0.09**	0.19 ± 0.10	0.14 ± 0.09
FUSE (CTRNN)	0.15 ± 0.02	0.18 ± 0.02	0.23 ± 0.07	0.18 ± 0.08	0.17 ± 0.09
FUSE (Neural-ODE)	0.15 ± 0.07	0.16 ± 0.06	0.25 ± 0.11	**0.17 ± 0.05**	**0.11 ± 0.05**

Table 3: MAE ($\downarrow$) of FUSE variants on the BH-rPPG dataset under controlled illumination levels (mean ± std).

Method	Low (8 lux)	Medium (42.4 lux)	High (104 lux)
FUSE (LTC)	0.31 ± 0.15	0.27 ± 0.06	0.26 ± 0.09
FUSE (CTRNN)	**0.26 ± 0.09**	**0.26 ± 0.06**	**0.21 ± 0.05**
FUSE (Neural-ODE)	0.38 ± 0.32	0.27 ± 0.08	0.24 ± 0.05

5.3 Explanation of ODE Variants

Table 4 analyzes different AC/DC temporal regularization designs. The proposed band-averaged harmonic constraint is $\mathcal{L}_{\text{band}}(x) = \frac{1}{K}\sum_{k=1}^{K} \phi\left(\frac{\partial^2 x}{\partial t^2} + \omega_k^2 x\right)$ where $\omega_k = 2\pi f_k$ spans a physiological cardiac band ($f_k \in [0.7, 3.0]$ Hz). This models AC components as distributed harmonic oscillators, allowing heart-rate variability.

Fixed-Frequency (ω_0)***:*** The variant $\frac{\partial^2 x}{\partial t^2} + \omega_0^2 x$ uses a single central frequency $\omega_0 = 2\pi f_0$ with $f_0 = 1.2$ Hz ($\omega_0 \approx 7.54$ rad/s). Removing frequency diversity significantly degrades performance (PURE: 0.25 MAE; BH-rPPG: 0.65 MAE), indicating that a single oscillator cannot capture physiological variability.
Generic Smoothness: Replacing the harmonic prior with $\|\partial_t^2 X\|_1$ enforces temporal smoothness without oscillatory structure. This improves over the fixed-frequency model (PURE: 0.21 MAE), but remains inferior to band-limited supervision, demonstrating that periodic structure not merely smoothness is essential.
Band and Channel Enforcement: Applying $\mathcal{L}_{\text{band}}^{AC}$ only to density ρ reduces error (0.17 MAE on PURE), while enforcing it jointly on (ρ, u, v) achieves the best results (0.15 / 0.28 / 0.57 MAE on PURE/BH-rPPG/VIPL-HR). This confirms that: (i) frequency-band modeling is critical, (ii) oscillator priors outperform generic second-order regularization, and (iii) multi-channel enforcement improves dynamical consistency.

Table 4: Ablation of AC/DC ODE design on PURE, BH-rPPG, and VIPL-HR. All variants use the same backbone and temporal model (LTC). Metrics: MAE (↓), RMSE (↓), and r (↑).

ODE Variant	PURE			BH-rPPG			VIPL-HR		
	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑
$\mathcal{L}_{\omega_0}^{AC} + \mathcal{L}_{\omega_0}^{DC}$	0.25	0.58	0.81	0.65	0.93	0.78	0.91	0.99	0.77
$\|\partial_t^2 X_{AC}\|_1 + \|\partial_t^2 X_{DC}\|_1$	0.21	0.50	0.86	0.48	0.72	0.84	0.78	0.90	0.83
$\mathcal{L}_{\text{band}}^{AC}(\rho) + \mathcal{L}_{\text{band}}^{DC}$	0.17	0.41	0.90	0.33	0.58	0.90	0.63	0.79	0.88
$\mathcal{L}_{\text{band}}^{AC}(\rho, u, v) + \mathcal{L}_{\text{band}}^{DC}$	**0.15**	**0.38**	**0.92**	**0.28**	**0.53**	**0.93**	**0.57**	**0.75**	**0.91**

5.4 Effect of PDE Transport Regularization

Table 5 evaluates the contribution of the proposed transport-based PDE constraints while keeping the backbone and temporal model (LTC) fixed.

Without PDE regularization, performance degrades noticeably (e.g., PURE: 0.20 MAE; BH-rPPG: 0.35 MAE). Introducing the advection residual $\mathcal{L}_{adv}$ improves spatio-temporal coherence, reducing error across datasets. Adding divergence regularization $\mathcal{L}_{div}$ further stabilizes the solution and improves correlation. Finally, incorporating the anti-collapse energy term $\mathcal{L}_{energy}$ yields the best performance (0.15 / 0.28 / 0.57 MAE on PURE/BH-rPPG/VIPL-HR).

The progressive improvement confirms that spatial transport modeling provides complementary gains beyond harmonic ODE supervision, and that preventing trivial density collapse is essential for stable pulsatile propagation.

5.5 Additional Experiments

Cross-Dataset Generalization: We evaluate transfer without target-domain fine-tuning across PURE (30 fps), VIPL-HR (25 fps), and BH-rPPG (15 fps). As shown in ***Supplementary Sec:1***, FUSE consistently surpasses ACuRE in all directions, with the largest gains under severe frame-rate mismatch (BH-rPPG),

Table 5: Ablation of the PDE transport components on PURE, BH-rPPG, and VIPL-HR. We progressively enable $\mathcal{L}_{adv}$, $\mathcal{L}_{div}$, and $\mathcal{L}_{energy}$. All results use the same backbone and temporal model (LTC); only PDE terms are varied. Metrics: MAE (↓), RMSE (↓), and r (↑).

PDE Components	PURE			BH-rPPG			VIPL-HR		
	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑	MAE ↓	RMSE ↓	r ↑
None (No PDE)	0.20	0.44	0.89	0.35	0.63	0.88	0.71	0.85	0.87
$\mathcal{L}_{adv}$	0.18	0.41	0.90	0.32	0.59	0.90	0.66	0.81	0.88
$\mathcal{L}_{adv} + \mathcal{L}_{div}$	0.16	0.39	0.91	0.30	0.56	0.91	0.61	0.78	0.89
$\mathcal{L}_{adv} + \mathcal{L}_{div} + \mathcal{L}_{energy}$	**0.15**	**0.38**	**0.92**	**0.28**	**0.53**	**0.93**	**0.57**	**0.75**	**0.91**

confirming improved invariance to temporal sampling and appearance shifts.

Temporal Modeling Comparison: Under identical backbone and regularization settings, we compare discrete recurrence (RNN/LSTM/GRU), TCN, and continuous-time variants. As reported in ***Supplementary Sec:2***, discrete models exhibit instability under frame-rate variation, whereas continuous-time architectures (Neural-ODE, CTRNN, LTC) achieve consistently lower error and higher correlation, validating continuous-time dynamics for physiological modeling.

Backbone Architecture Ablation: Using a fixed LTC temporal module, we evaluate multiple spatial backbones. As detailed in ***Supplementary Sec:3***, 3D backbones yield the lowest absolute errors, while PhysNet attains competitive accuracy with only 2.53M parameters (vs. 66.5M for ResNet3D-18 and 39.5M for ViT3D), demonstrating strong parameter efficiency and backbone-agnostic compatibility of the proposed regularization.

Parameter Sensitivity: We analyze the effect of the harmonic weight α with fixed PDE coefficients. As shown in ***Supplementary Sec:4***, both under- and over-weighting degrade performance, while $\alpha = 0.1$ provides the best balance between harmonic structure and transport regularization across datasets.

6 Conclusion and Future Work

We presented FUSE, a filter-free framework for camera-based SpO_2 estimation that replaces frequency-domain signal decomposition with differential constraints. The proposed formulation models pulsatile dynamics through harmonic ODE regularization while enforcing spatial-temporal coherence via transport-based PDE constraints. Combined with continuous-time dynamical modeling, this approach enables robust AC/DC representation learning without relying on noisy frequency filtering. Extensive experiments demonstrate improved stability and cross-dataset generalization under varying frame rates and visual conditions.

Future work will explore adaptive physiological priors for personalized modeling, extend the framework to multi-signal physiological estimation (e.g., HR and respiration), and investigate broader physics-informed formulations for remote health monitoring from video.

Acknowledgments

This research is funded by the Babysensor AS , Norway and the Østfold University of Applied Sciences.

References

1. Agarwal, V., Saikia, T., Gupta, A.K., Gupta, P.: Shine: Synergizing transformers with contrastive learning for thriving rppg-based spo2 estimation. Expert Systems with Applications p. 129190 (2025)
2. Ahmad, S., Mishra, D., Bano, S., Chanda, S., Rawat, Y.S.: Acure: Accurate continuity-regularized spo2 estimation using liquid time-constant networks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7250–7259 (March 2026)
3. Ahmad, S., Mukherjee, S., Chanda, S., Palaiahnakote, S., Pal, U., Pedersen, M.: Phase: Physiological dynamics-based attention for spo 2 estimation. In: Scandinavian Conference on Image Analysis. pp. 30–44. Springer (2025)
4. Akamatsu, Y., Onishi, Y., Imaoka, H.: 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). pp. 1–5. IEEE (2023)
5. Bowes Jr III, W.A., Corke, B.C., Hulka, J.: Pulse oximetry: a review of the theory, accuracy, and clinical applications. Obstetrics & Gynecology **74**(3 Part 2), 541–546 (1989)
6. Centers for Disease Control and Prevention: Healthcare-associated infections (hais). `https://www.cdc.gov/healthcare-associated-infections/index.html` (2023), accessed: June 29, 2026
7. Chen, R.T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.: Neural ordinary differential equations. Advances in Neural Information Processing Systems **31**, 6571–6583 (2018)
8. Cheng, C.H., Yuen, Z., Chen, S., Wong, K.L., Chin, J.W., Chan, T.T., So, R.H.: Contactless blood oxygen saturation estimation from facial videos using deep learning. Bioengineering **11**(3), 251 (2024)
9. Cho, K., Van Merriënboer, B., Gulçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1724–1734 (2014)
10. Chowdhury, M.H., Reaz, M.B.I., Ali, S.H.M., Khan, M.S., Chowdhury, M.E.: Rosenet: Leveraging remote photoplethysmography to estimate oxygen saturation using deep learning. Biomedical Signal Processing and Control **100**, 107105 (2025)
11. Couzin-Frankel, J.: The mystery of the pandemic's 'happy hypoxia'. Science **368**(6499), 455–456 (2020)
12. Funahashi, K.i., Nakamura, Y.: Approximation of dynamical systems by continuous time recurrent neural networks. Neural networks **6**(6), 801–806 (1993)
13. Hasani, R., Lechner, M., Amini, A., Rus, D., Grosu, R.: Liquid time-constant networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 7657–7666 (2021)
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

15. Hu, M., Wu, X., Wang, X., Xing, Y., An, N., Shi, P.: Contactless blood oxygen estimation from face videos: A multi-model fusion method based on deep learning. Biomedical Signal Processing and Control **81**, 104487 (2023)
16. Kong, L., Zhao, Y., Dong, L., Jian, Y., Jin, X., Li, B., Feng, Y., Liu, M., Liu, X., Wu, H.: Non-contact detection of oxygen saturation based on visible light imaging device using ambient light. Optics express **21**(15), 17464–17471 (2013)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
18. Mathew, J., Tian, X., Wong, C.W., Ho, S., Milton, D.K., Wu, M.: Remote blood oxygen estimation from videos using neural networks. IEEE journal of biomedical and health informatics **27**(8), 3710–3720 (2023)
19. McKinsey Company: Telehealth: A quarter-trillion-dollar post-covid-19 reality? `https://www.mckinsey.com/industries/healthcare/our-insights/telehealth-a-quarter-trillion-dollar-post-covid-19-reality` (2021), accessed: June 29, 2026
20. Niu, X., Han, H., Shan, S., Chen, X.: Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14. pp. 562–576. Springer (2019)
21. Peng, J., Su, W., Chen, H., Sun, J., Tian, Z.: Cl-spo2net: Contrastive learning spatiotemporal attention network for non-contact video-based spo2 estimation. Bioengineering **11**(2), 113 (2024)
22. Rahman, H., Ahmed, M.U., Begum, S.: Non-contact physiological parameters extraction using facial video considering illumination, motion, movement and vibration. IEEE Transactions on Biomedical Engineering **67**(1), 88–98 (2019)
23. Shao, J., Lin, L., Niwayama, M., Kudo, N., Yamamoto, K.: Theoretical and experimental studies on linear and nonlinear algorithms for the measurement of muscle oxygenation using continuous-wave near-infrared spectroscopy. Optical Engineering **40**(10), 2293–2301 (2001)
24. Stam, N., Rebollo, D., Asemu, Y.M., Aklu, L., et al.: Pulse oximetry in low-resource settings during the covid-19 pandemic. The Lancet Global Health **8**(9), e1121–e1122 (2020)
25. Stricker, R., Müller, S., Gross, H.M.: Non-contact video-based pulse rate measurement on a mobile service robot. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication. pp. 1056–1062. IEEE (2014)
26. Sun, X., Wen, T., Chen, W., Huang, B.: Ccspo 2 net: Camera-based contactless oxygen saturation measurement foundation model in clinical settings. IEEE Transactions on Instrumentation and Measurement **73**, 1–11 (2024)
27. Tamura, T.: Current progress of photoplethysmography and spo2 for health monitoring. Biomedical engineering letters **9**(1), 21–36 (2019)
28. Tarassenko, L., Villarroel, M., Guazzi, A., Jorge, J., et al.: Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. Physiological Measurement **35**(5), 807–831 (2014)
29. World Health Organization: Chronic obstructive pulmonary disease (copd). `https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)` (2024), accessed: June 29, 2026
30. Yang, Z., Wang, H., Lu, F.: 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)

31. Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In: Proceedings of the British Machine Vision Conference (BMVC). pp. 1–12 (2019)
32. Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit tadas baltrušaitis. In: IEEE International Conference on Automatic Face and Gesture Recognition (2018)
33. Zhu, S., Liu, S., Jing, X., Yang, Y., She, C.: Innovative approaches in imaging photoplethysmography for remote blood oxygen monitoring. Scientific Reports **14**(1), 19144 (2024)