Vital Signs Estimation Using Remote Photoplethysmography rPPG

by

Mohamed Khalil BEN SALAH

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, DECEMBER 1, 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mrs. Rita Noumeir, Thesis supervisor
Department of Electrical Engineering, École de Technologie Supérieure

Mr. Mr. Philippe Jouvet, Thesis Co-Supervisor
Pediatric Intensivist - Ste. Justine Hospital Montréal, Université de Montréal

Mr. Marco Pedersoli, Chair, Board of Examiners
Department of Systems Engineering, École de Technologie Supérieure

Mr. Jean-Marc Lina, Member of the Jury
Department of Electrical Engineering, École de Technologie Supérieure

Mr. Benjamin de Leener, External Independent Examiner
Department of Computer Engineering and Software Engineering, Polytechnique Montréal

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON NOVEMBER 24, 2025

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to begin by expressing my sincere gratitude to my supervisor, Professor Rita Noumeir, for her guidance, patience, and unwavering support throughout my doctoral research. Her structured approach, attention to detail, and commitment to scientific rigor have significantly shaped the quality of this work and my development as a researcher. I am thankful for her availability and encouragement during every stage of this thesis.

I am equally grateful to Dr. Philippe Jouvet for his clinical expertise, thoughtful feedback, and active involvement in our research discussions. His efforts to maintain a consistent connection between the clinical and engineering teams helped ground this work in real-world challenges and ensured that our contributions remained relevant to the PICU context.

I would also like to thank the members of the examination committee for taking the time to review this thesis and for providing constructive feedback that strengthened the final version.

To my colleagues at the Biomedical Information Processing Laboratory (LATIS) and CHU Sainte-Justine, thank you for the collaborative spirit, technical discussions, and support throughout this journey. Monthly interdisciplinary meetings allowed us to identify challenges early, share resources, and maintain momentum across projects. I am also grateful to the LATIS group at ÉTS for creating an environment that fostered focused research, peer support, and meaningful exchanges.

To my family in Tunisia, thank you for your constant encouragement, patience, and belief in me throughout the years, despite the distance. I would like to express my heartfelt gratitude to my mother for her constant encouragement and strength throughout this journey. Her values and resilience have shaped my path more than words can express. To my father, thank you for your continued support, your presence in every milestone, and your belief in my work. This accomplishment would not have been possible without both of you.

Estimation des signes vitaux à l’aide photopléthysmographie à distance rPPG

Mohamed Khalil BEN SALAH

RÉSUMÉ

La surveillance des signes vitaux dans les Unités de Soins Intensifs Pédiatriques (USIP) est essentielle pour la prise en charge des patients pédiatriques vulnérables. Les approches conventionnelles, telles que l’électrocardiographie, reposent sur un contact physique et sont souvent invasives, coûteuses et inadaptées aux nouveau-nés ou aux patients atteints de maladies contagieuses. La photopléthysmographie à distance (rPPG) offre une alternative sans contact en capturant les variations subtiles de la couleur de la peau causées par le flux sanguin pulsatile. En soins intensifs pédiatriques, elle constitue une solution plus sûre que les capteurs adhésifs, qui peuvent provoquer des irritations et augmenter le risque d’infection. Cependant, le déploiement de la rPPG dans des environnements cliniques réels reste difficile en raison des occlusions fréquentes dues à l’équipement médical, des mouvements des patients, de la variabilité de l’éclairage et d’un décalage de domaine entre les données de laboratoire contrôlées et les enregistrements en USIP. Ces limitations sont aggravées par la rareté des ensembles de données cliniques annotées. Pour surmonter ces contraintes, il est nécessaire de développer des modèles physiologiquement interprétables, efficaces sur le plan computationnel et résilients aux changements de domaine. Cette thèse présente un cadre unifié qui intègre l’apprentissage de caractéristiques spatiotemporelles efficaces, la détection de régions anatomiquement cohérentes et un pré-entraînement auto-supervisé basé sur un curriculum pour obtenir une estimation précise et en temps réel de la fréquence cardiaque dans des environnements cliniques complexes.

Pour extraire des signaux rPPG fiables à partir de vidéos faciales non contraintes, une architecture hybride est proposée, combinant des blocs convolutionnels 3D avec des noyaux de différence temporelle (3DCDC-T) et une auto-attention multi-têtes issue des transformateurs de vision. Le modèle capture les gradients spatiotemporels locaux indicatifs des variations du volume sanguin tout en modélisant les dépendances à plus longue portée nécessaires pour résoudre les cycles cardiaques complets. Les mécanismes d’attention affinent davantage la focalisation des caractéristiques sur les régions faciales physiologiquement informatives, et la conception feed-forward garantit l’efficacité computationnelle en limitant l’entrée du transformateur à des plongements de caractéristiques compacts. Évalué sur des ensembles de données publics, le modèle atteint une MAE de 0,79 bpm et une RMSE de 0,80 bpm, avec une corrélation de Pearson de 0,99, améliorant ainsi les méthodes existantes tant en termes de précision que de coût d’inférence.

Une estimation précise de la rPPG nécessite un suivi anatomique stable des régions du visage et thoraco-abdominales, en particulier dans les vidéos affectées par la rotation, l’inclinaison du lit ou l’occlusion par le personnel soignant. Un module de détection dédié est développé en utilisant le modèle Divided Space–Time Mamba (DST-Mamba). Cette architecture découple le traitement spatial et temporel grâce à des Modèles d’Espace d’État Sélectifs (SSM), permettant une complexité en temps linéaire et une inférence à faible latence sur des séquences vidéo plus longues. Le modèle prédit des boîtes englobantes orientées (OBB) pour préserver l'alignement en rotation sous des angles de caméra non standards et intègre des entrées RGB-D pour améliorer la robustesse face aux occlusions visuelles. DST-Mamba atteint 0,96 mAP@0,5 et 0,95 d'IoU tournée sur un ensemble de données cliniques, maintenant la stabilité temporelle tout en fonctionnant à 23 FPS sur du matériel standard.

Pour pallier la rareté des données étiquetées en USIP, une stratégie d'apprentissage auto-supervisé basée sur un curriculum est introduite. Un contrôleur de masquage adaptatif basé sur Mamba attribue des scores d'importance spatiotemporelle aux patchs d'entrée et applique un masquage stratégique en utilisant l'échantillonnage de Gumbel différentiable. Ce masquage contradictoire force le modèle à reconstruire les signaux physiologiques à partir d'entrées dégradées, favorisant ainsi la robustesse aux occlusions et distractions cliniques. Le processus d'apprentissage suit un curriculum structuré : entraînement initial sur des ensembles de données publics, simulation de motifs d'occlusion observés dans les enregistrements de l'USIP, et adaptation de domaine sur 500 vidéos cliniques non étiquetées. Un module de distillation enseignant-élève léger transfère les a priori physiologiques de modèles experts. Ce pipeline réduit les besoins en données supervisées de 80 %, atteignant une MAE de 3,2 bpm en utilisant seulement 160 patients étiquetés, contre 18,2 bpm avec un entraînement supervisé direct.

Le cadre est validé sur un vaste ensemble de données collectées au CHU Sainte-Justine, démontrant une généralisation à travers les âges, les teintes de peau et les conditions d'occlusion. Le système maintient une MAE inférieure à 7,2 bpm avec plus de 70% d'occlusion faciale, atteignant 3,8 bpm pour les nouveau-nés et 3,5 bpm pour les patients sous ventilation mécanique. Il fonctionne en temps réel dans les contraintes cliniques, consommant 169,7 GFLOPs et 6,1 Go de mémoire avec un débit de 30 FPS. Ensemble, ces contributions lèvent les obstacles clés au déploiement clinique de la rPPG, notamment l'adaptation de domaine, le suivi anatomique et l'efficacité des données, faisant progresser la surveillance physiologique sans contact vers une utilisation pratique en soins intensifs pédiatriques.

Mots-clés: unité de soins intensifs pédiatriques, photopléthysmographie à distance, surveillance sans contact, apprentissage auto-supervisé, architecture mamba, transformateurs de vision

Vital Signs Estimation Using Remote Photoplethysmography rPPG

Mohamed Khalil BEN SALAH

ABSTRACT

Vital sign monitoring in Pediatric Intensive Care Units (PICUs) is critical for managing vulnerable pediatric patients. Conventional approaches, such as electrocardiography, rely on physical contact and are often invasive, costly, and unsuitable for newborns or patients with contagious conditions. Remote photoplethysmography (rPPG) offers a non-contact alternative by capturing subtle variations in skin color caused by pulsatile blood flow. In pediatric intensive care, it provides a safer solution than adhesive sensors, which can cause irritation and increase the risk of infection. However, deploying rPPG in real clinical environments remains challenging due to frequent occlusions from medical equipment, patient motion, illumination variability, and a domain gap between controlled laboratory data and PICU recordings. These limitations are compounded by the scarcity of annotated clinical datasets. Addressing these constraints requires models that are physiologically interpretable, computationally efficient, and resilient to domain shifts. This thesis introduces a unified framework that integrates efficient spatiotemporal feature learning, anatomically consistent region detection, and curriculum-based self-supervised pretraining to achieve accurate and real-time estimation of heart rate in complex clinical environments.

To extract reliable rPPG signals from unconstrained facial videos, a hybrid architecture is proposed that combines 3D convolutional blocks with temporal difference kernels (3DCDC-T) and multi-head self-attention from vision transformers. The model captures local spatiotemporal gradients indicative of blood volume changes while modeling longer-range dependencies required to resolve complete cardiac cycles. Attention mechanisms further refine feature focus on physiologically informative facial regions, and the feed-forward design ensures computational efficiency by limiting the transformer's input to compact feature embeddings. Evaluated on public datasets, the model achieves an MAE of 0.79 bpm and RMSE of 0.80 bpm, with a Pearson correlation of 0.99, improving over existing methods both in accuracy and inference cost.

Accurate rPPG estimation requires stable anatomical tracking of face and thoracoabdominal regions, particularly in videos affected by rotation, bed tilt, or caregiver occlusion. A dedicated detection module is developed using the Divided Space–Time Mamba (DST-Mamba) model. This architecture decouples spatial and temporal processing through Selective State Space Models (SSMs), enabling linear-time complexity and low-latency inference across longer video sequences. The model predicts oriented bounding boxes (OBBs) to preserve rotation alignment under non-standard camera angles and integrates RGB-D inputs to improve robustness against visual occlusions. DST-Mamba achieves 0.96 mAP@0.5 and 0.95 rotated IoU on a clinical dataset, maintaining temporal stability while operating at 23 FPS on standard hardware.

To mitigate the scarcity of labeled PICU data, a curriculum-based self-supervised learning strategy is introduced. A Mamba-based adaptive masking controller assigns spatiotemporal importance scores to input patches and applies strategic masking using differentiable Gumbel sampling. This adversarial masking forces the model to reconstruct physiological signals from degraded inputs, encouraging robustness to clinical occlusions and distractions. The learning process follows a structured curriculum: initial training on public datasets, simulation of occlusion patterns observed in PICU recordings, and domain adaptation on 500 unlabeled clinical videos. A lightweight teacher–student distillation module transfers physiological priors from expert models. This pipeline reduces supervised data requirements by 80%, achieving an MAE of 3.2 bpm using only 160 labeled patients, compared to 18.2 bpm with direct supervised training.

The framework is validated on an extensive dataset collected at CHU Sainte-Justine, demonstrating generalization across ages, skin tones, and occlusion conditions. The system maintains MAE under 7.2 bpm with over 70% facial occlusion, achieving 3.8 bpm for neonates and 3.5 bpm for mechanically ventilated patients. It operates in real-time within clinical constraints, consuming 169.7 GFLOPs and 6.1 GB memory at 30 FPS throughput. Together, these contributions address key barriers to clinical rPPG deployment, including domain adaptation, anatomical tracking, and data efficiency, moving non-contact physiological monitoring toward practical use in pediatric intensive care.

Keywords: pediatric Intensive Care Unit, remote photoplethysmography, non-contact monitoring, self-supervised learning, mamba architecture, vision transformers

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

3D-CNNs	3D-Convolutional Neural Networks
3DCDC-T	3D Central Difference Convolutional (Temporal)
AI	Artificial Intelligence
AMN	Adaptive Masking Network
ARDS	Acute Respiratory Distress Syndrome
bpm	Beats Per Minute
BSS	Blind Source Separation
BVP	Blood Volume Pulse
CBAM	Convolutional Block Attention Module
CHROM	Chrominance-based method
CHUSJ	CHU Sainte-Justine Research Center
CNNs	Convolutional Neural Networks
COCO	Common Objects in Context
CSL	Circular Spatial Layout
DETR	DEtection TRansformer
DST-Mamba	Divided Space-Time Mamba
ECG	Electrocardiogram
EVM	Eulerian Video Magnification
FDA	Food and Drug Administration

GRU	Gated Recurrent Unit
HR	Heart Rate
ICA	Independent Component Analysis
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MHSA	Multi-Head Self-Attention
MPE	Mean Percentage Error
NICU	Neonatal Intensive Care Units
OBB	Oriented Bounding Boxes
PCCPs	Predetermined Change Control Plans
PEWS	Pediatric Early Warning Score
PICU	Pediatric Intensive Care Units
PMA	Premarket Approval
POS	Plane-Orthogonal-to-Skin
R	Pearson’s Correlation Coefficient
RAdam	Rectified Adam
rIoU	Rotated Intersection over Union
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
ROI	Regions of Interest

rPPG	Remote Photoplethysmography
RR	Respiratory Rate
S4	Structured State Space Sequence Model
SaMD	Software as a Medical Device
SNR	Signal-to-Noise Ratio
SpO_2	Oxygen Saturation
SSL	Self-Supervised Learning
SSM	Selective State Space Models
TDC	Temporal Difference Convolution
TSM	Temporal Shift Module
UMT	Unmasked Teacher
USIP	Unités de Soins Intensifs Pédiatriques
VideoMAE	Video Masked Autoencoders
ViTs	Vision Transformers
ViViT	Video Vision Transformer

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

bpm	Beats Per Minute
fps	Frames Per Second
GB	Gigabytes
GFLOPs	Giga Floating-Point Operations Per Second
Hz	Hertz
ms	Milliseconds

INTRODUCTION

Monitoring vital signs in Pediatric Intensive Care Units (PICUs) is essential for ensuring patient safety and enabling timely clinical interventions, particularly given the fragile health conditions of pediatric patients. Providing effective care in the PICU relies on continuously monitoring essential physiological parameters to capture early signs of deterioration. Vital signs provide clinical staff with crucial insights into the patient's response to treatment and alert them to potential complications. Among the primary vital signs of interest in this setting are Heart Rate (HR), Respiratory Rate (RR), and Oxygen Saturation (SpO_2). Early detection of instability such as tachycardia, hypoxia, or respiratory distress requires uninterrupted signals with sufficient temporal resolution to capture the rapid physiological changes typical of critically ill infants and children.

Traditional contact-based sensors, including electrocardiograms (ECGs), pulse oximeters, and wearable monitoring devices, remain the clinical standard for measuring heart rate, oxygen saturation, and respiratory rate. However, in the PICU, their use presents multiple challenges. Adhesive electrodes can cause skin irritation, increase the risk of infection, and create discomfort, particularly in neonates with immature skin. Wearable devices may restrict patient mobility, interfere with routine caregiving procedures, and require frequent maintenance. In addition to the physical burden, these systems often deliver only short-term recordings and may fail to capture transient abnormalities. Furthermore, in the context of infection control or highly contagious diseases, contact-based monitoring becomes impractical. Setup complexity and cost considerations further limit scalability and usability in fast-paced or resource-constrained environments. These limitations underscore the urgent need for non-contact alternatives better suited to the pediatric critical care context.

Recent studies have demonstrated the feasibility of remotely estimating heart rate using remote photoplethysmography (rPPG). This technique analyzes skin reflectance variations caused by cardiac cycle captured in video recordings to extract blood volume pulse information. rPPG provides a non-contact solution for monitoring vital signs, particularly in cases where physical contact is challenging or undesirable. The use of RGB cameras to capture physiological signals from skin pixel intensity has been shown to enable reliable estimation of HR and HR variability. By analyzing the periodicity of pixel intensity variations over time, especially in RGB or thermal modalities, it is possible to reconstruct waveforms that correlate with physiological signals. This approach is particularly suitable for pediatric patients, including neonates with fragile skin, individuals with dermatological conditions, or patients under infection control measures such as isolation during the COVID-19 pandemic.

However, most of these techniques are developed and tested in controlled laboratory settings, often using adult participants under fixed lighting and at limited distances from the camera. These assumptions do not hold in real clinical settings. The PICU environment presents a particularly challenging set of conditions for rPPG deployment. Patients are frequently surrounded by obstructive medical devices such as ventilator tubing, nasal cannulas, oxygen masks, and medical wires, all of which occlude parts of the face. Lighting conditions in the PICU vary significantly due to clinical routines, day-night cycles, and mobile equipment. Furthermore, spontaneous movements, changes in posture, and variations in caregiver handling introduce noise and motion artifacts. Pediatric patients also differ physiologically from adults, not only in terms of vital sign ranges but also in terms of facial morphology and skin properties. These challenges severely limit the generalization of rPPG models trained on laboratory datasets to clinical pediatric applications.

The deployment of reliable rPPG monitoring in PICUs faces three fundamental challenges that existing approaches have not resolved. First, there is a severe domain gap between laboratory training data and clinical deployment environments, which regularly causes model failure. Public datasets such as UBFC-rPPG and PURE contain high-quality recordings collected under controlled conditions, but they do not capture the visual complexity, occlusions, and physiological variability characteristic of PICU settings. Models trained on these datasets fail to generalize well to clinical data, with performance declining substantially. Second, the scarcity of annotated pediatric video with synchronized ground truth prevents supervised learning from reaching clinical-grade accuracy. Privacy regulations, ethical constraints, and the technical burden of aligning video with gold-standard measurements restrict dataset scale. The CHU Sainte-Justine Research Center (CHUSJ) corpus of five hundred pediatric patients is among the largest available, yet it remains insufficient to train deep models from scratch using conventional supervision. Third, current architectures fail to balance computational efficiency with robustness to clinical artifacts. Transformer-based designs achieve strong accuracy but require resources that exceed bedside workstation capabilities, whereas efficient CNN-based designs lack the temporal modeling capacity needed to maintain stable measurements under motion and occlusion.

The primary objective of this thesis is to develop a robust, non-contact physiological monitoring framework capable of estimating heart rate with clinical-grade accuracy in the Pediatric Intensive Care Unit (PICU). Achieving this goal requires overcoming three interconnected barriers: the efficient modeling of spatiotemporal physiological signals, the stable tracking of anatomical regions under occlusion, and the scarcity of labeled clinical data.

The first challenge lies in the architectural limitations of current deep learning models. Existing approaches for rPPG face a fundamental trade-off: 3D-CNNs possess a strong inductive bias for capturing local skin color variations but struggle to model long-range temporal dependencies due to their limited receptive fields. Conversely, Vision Transformers excel at modeling global context but lack local inductive bias and are often computationally prohibitive for real-time applications. We hypothesize that a hybrid architecture, which sequentially integrates 3D-CNNs to capture local temporal gradients and Vision Transformers to model global context in a feed-forward manner, will outperform parallel or single-architecture models. This sequential refinement is expected to enhance signal fidelity by leveraging the specific strengths of both architectures, combining the local inductive bias of convolutions with the global model capacity of attention mechanisms.

Even with a high-fidelity signal estimation model, reliable monitoring is contingent upon the stable tracking of regions of interest (ROI). However, the PICU environment presents unique tracking challenges, including patient rotation, bed tilt, and frequent occlusions. Standard frame-based detectors suffer from temporal instability, while video transformers that model temporal consistency inherently suffer from quadratic complexity ($O(N^2)$), rendering them unsuitable for real-time bedside monitoring. We hypothesize that decoupling spatial and temporal modeling using Selective State Space Models (Mamba) will yield detection accuracy and temporal stability comparable to transformer-based methods, but with linear-time complexity ($O(N)$). By factorizing the learning process, the model should allow for the robust tracking of oriented bounding boxes under occlusion while strictly adhering to clinical latency constraints.

Finally, the deployment of these models is hindered by a significant domain gap between controlled laboratory datasets and the noisy, occluded environment of the PICU. Compounding this issue is the scarcity of labeled pediatric data, which limits the efficacy of supervised learning. Furthermore, standard self-supervised methods, such as random masking, fail to account for the specific, structured occlusions (e.g., oxygen masks, tubes) found in clinical settings. We hypothesize that an adaptive masking strategy, which adversarially targets and obscures signal-rich spatiotemporal regions, will force the model to learn more robust and redundant physiological features. When implemented within a progressive curriculum, moving from clean data to synthetic occlusions and finally to real clinical data, this approach is expected to significantly reduce the reliance on labeled data while improving generalization to the complex clinical domain.

The research began by addressing the core task of rPPG signal extraction. In my first paper, "rPPG Estimation: Vision Transformer With 3-D Temporal Central Difference", I developed a novel hybrid architecture that combined 3D-Convolutional Neural Networks (3D-CNNs) with a Video Vision Transformer (ViViT). To this end, I designed a novel hybrid video architectures that combine the local inductive biases of 3D convolutional operators with the global temporal context modeled by video vision transformers. This model focused on capturing fine-scale pulsatile dynamics and integrate them with attention mechanisms that maintain long-range coherence. This feed-forward approach demonstrated superior performance on public datasets, establishing a powerful baseline for accurate rPPG estimation. For clinicians, this enables reliable HR estimation without contact; for technicians, it demonstrates sequential refinement as a scalable alternative to parallel architectures.

While the first model was effective, its application in the PICU was predicated on a critical assumption: that the anatomical regions of interest (the face and thoracoabdominal area) could be reliably located. In the cluttered PICU environment, this is a non-trivial problem. The second contribution introduces Divided Space-Time Mamba, a self-supervised architecture using State Space Models for linear-time complexity "PICU Face and Thoracoabdominal Detection using Self-Supervised Divided Space-Time Mamba". I introduced an efficient detector based on State Space Models (Mamba) , which have linear-time complexity, making them suitable for real-time clinical hardware. The aim is temporally stable, orientation-aware localization of the face and thoracoabdominal regions using oriented bounding boxes that remain consistent under pose changes, caregiver interactions, and device occlusions. To overcome the profound lack of annotated PICU data, I leveraged self-supervised pre-training on over 50,000 domain-specific video clips, enabling the model to learn robust features directly from the target environment. This innovation addresses data scarcity and temporal instability, providing clinicians with jitter-free tracking for vital monitoring and technical experts with an efficient alternative to quadratic-attention transformers.

Finally, to optimize rPPG for PICUs and enhance generalization "Adaptive Video Masking with Curriculum-based Self-Supervised Learning for rPPG Estimation in PICU", an adaptive video masking framework with curriculum-based SSL was developed. This work introduces a complete framework tailored for the PICU, using an efficient architecture. The core innovations of this work lie in two complementary components. First, I introduce a progressive curriculum learning strategy that gradually increases the training difficulty, beginning with clean, high-quality lab recordings, followed by videos augmented with synthetic occlusions, and culminating in adaptation to a large-scale, unlabeled dataset composed of 500 PICU patient videos. This staged approach systematically bridges the domain gap between controlled environments and the complex clinical setting. Second, I replace conventional random masking with a learnable, Mamba-based masking controller that adversarially suppresses the most informative spatiotemporal regions. This adaptive mechanism compels the model to learn physiologically meaningful and spatially redundant representations, thereby enhancing robustness to real-world occlusions without requiring explicit region-of-interest (ROI) detection or face tracking.

This dissertation is structured as an article-based thesis that develops the technical approach in a progressive manner from foundations to deployment. Chapter 1 surveys the state of the art in camera-based physiological monitoring, covering classical signal processing pipelines and recent deep learning methods. The review analyzes why techniques validated on adult laboratory recordings often fail in pediatric intensive care, and it identifies the specific gaps that motivate our architectural and training choices in subsequent chapters.

Chapter 2 presents the first article, “rPPG Estimation: Vision Transformer With 3-D Temporal Central Difference” (IEEE Transactions on Instrumentation and Measurement, 2025). This study establishes the baseline hybrid design that couples temporal-difference 3D operators with attention for long-range context, and it demonstrates consistent gains on standard rPPG benchmarks relative to prior models.

Chapter 3 introduces the second article, “Self-Supervised Divided Space-Time Mamba for PICU Face and Thoracoabdominal Detection.” Here, a selective state-space formulation separates spatial and temporal processing to achieve linear-time complexity while maintaining orientation-aware, temporally stable detection under clinical occlusions and motion. This chapter connects detection stability to downstream physiological estimation by quantifying jitter and rotated overlap over time.

Chapter 4 develops the third article, “Non-Contact Physiological Monitoring in Pediatric Intensive Care Units via Adaptive Masking and Self-Supervised Learning.” This work integrates the architectural elements with a training curriculum that moves from laboratory videos to synthetic occlusions and then to hospital data. Adaptive masking and teacher–student distillation are used to learn representations that preserve pulsatile signal fidelity in the presence of occlusion, illumination change, and caregiver interaction.

The thesis concludes with Chapter 5, which synthesizes findings across the three articles, discusses limitations, and outlines directions for clinical translation, and provides the overall conclusions. Appendices report extended experiments, dataset and ethics details, and implementation notes to support reproducibility.

CHAPTER 1

LITERATURE REVIEW

In this chapter, we critically examine the existing literature on non-contact vital sign monitoring, with a specific focus on its deployment in pediatric intensive care units (PICUs). The review draws from clinical, engineering, and ethical perspectives to provide a comprehensive foundation for the thesis. Emphasis is placed on the persistent challenges of translating controlled laboratory advancements into real-world clinical settings, where patient fragility, environmental variability, and strict data governance impose constraints on system design and deployment. From a clinical standpoint, this necessitates monitoring tools that minimize patient disturbance while maintaining accuracy and reliability. From a technical perspective, it highlights the importance of developing models that are both computationally efficient and capable of generalizing across heterogeneous patient populations and recording conditions.

The chapter begins by outlining the clinical motivations for non-contact monitoring in the PICU and traces the evolution of relevant technologies, identifying key limitations that remain unresolved. While previous reviews have largely concentrated on general rPPG methodologies, our analysis expands the scope to include underexplored areas such as multimodal data fusion, domain adaptation, and the ethical implications of AI-driven monitoring in pediatric care. Within this context, we position our own contributions as targeted responses to these challenges, specifically through the development of hybrid model architectures and self-supervised learning strategies designed to improve clinical utility and robustness. Anchored by these clinical priorities, the subsequent section addresses the specific demands and constraints that define vital sign monitoring in the PICU environment.

1.1 Foundations of Non-Contact Vital Sign Monitoring

Continuous monitoring of vital signs in the Pediatric Intensive Care Unit (PICU) is a cornerstone of clinical care, enabling early detection of physiological instability and guiding timely interventions in fragile patients (Ruhrberg Estévez *et al.*, 2025). Standard monitoring tools, including electrocardiogram (ECG) electrodes and pulse oximeters, depend on adhesive contact with the skin, which poses substantial challenges in neonatal and pediatric populations. These sensors often irritate delicate skin, elevate infection risk, and interfere with caregiving by restricting patient mobility and requiring frequent repositioning (Senechal *et al.*, 2023; Krbec *et al.*, 2024). In addition, the physical presence of wires and probes complicates routine handling and can interrupt continuity of care. Non-contact alternatives, such as remote photoplethysmography (rPPG), offer the potential to extract heart rate and respiratory metrics directly from video recordings, eliminating the burden of physical sensors (Kumar, Veeraraghavan & Sabharwal, 2015).

1.1.1 Evolution of Pediatric Critical Care Monitoring

The evolution of vital sign monitoring in pediatric intensive care reflects a shift from intermittent manual assessments to continuous, technology-enabled surveillance. Early monitoring practices in pediatric settings were based on nursing observations such as pulse palpation, respiratory rate counting, and mercury-based thermometry. While non-invasive, these methods provided only limited temporal snapshots of the patient's physiological state (Sanchez-Pinto, Venable, Fahrenbach & Churpek, 2018). The adoption of continuous electrocardiography (ECG) in the 1960s marked a significant advancement, offering real-time cardiac monitoring capabilities. However, this progress introduced a number of unresolved limitations. Adhesive electrodes often lead to skin injuries, particularly in premature neonates; wire-based systems restrict mobility and interfere with parent-infant bonding; and the presence of multiple sensors and leads increases the risk of infection.

The pediatric population poses distinct challenges that are not adequately addressed by monitoring technologies designed for adults. Neonatal skin, being thinner and more permeable, is especially vulnerable to damage during adhesive removal. Reports indicate that a high proportion of extremely low birth weight infants suffer from adhesive-related skin injuries (McNichol, Lund, Rosen & Gray, 2013). Additionally, the limited body surface area in neonates constrains the placement of sensors, and their elevated heart and respiratory rates necessitate higher sampling frequencies to ensure accurate signal acquisition. These considerations underscore the need for pediatric-specific monitoring solutions developed with a clear understanding of neonatal physiology, rather than adaptations of adult-oriented systems.

1.1.2 Principles of PPG and rPPG

Photoplethysmography (PPG) is an optical technique for measuring blood volume pulse by detecting variations in light absorption and reflection within vascularized tissue (Allen, 2007). When illuminated by a light source, such as an LED, biological tissue absorbs and scatters part of the incident light, while a photodetector captures the portion that is reflected or transmitted. The pulsatile component of this signal corresponds to the rhythmic changes in arterial blood volume driven by the cardiac cycle, whereas the slowly varying baseline reflects non-pulsatile tissue and venous blood. In practice, PPG waveforms provide valuable information on heart rate, oxygen saturation, and respiratory-induced modulations (Park, Seok, Kim & Shin, 2022).

Beyond its acquisition principle, the morphology of the PPG waveform itself is physiologically meaningful. Each cardiac cycle typically produces a systolic upstroke culminating in a sharp peak, corresponding to the rapid ejection of blood from the left ventricle. Following this peak, a secondary feature known as the dicrotic notch reflects the closure of the aortic valve, and a smaller diastolic wave may be observed thereafter. The amplitude, timing, and relative prominence of these features can be influenced by vascular tone, arterial stiffness, and respiratory activity, making PPG an indirect yet powerful tool for assessing cardiovascular dynamics (Liang, Chen, Ward & Elgendi, 2018; Moraes *et al.*, 2018).

Traditional PPG is typically acquired via contact-based probes, such as finger clips or adhesive sensors, which ensure consistent illumination and detection geometry. However, these devices have limitations in neonatal and pediatric care due to their reliance on direct skin contact, the risk of irritation, and their susceptibility to motion artifacts or sensor displacement.

Remote photoplethysmography (rPPG) extends these principles to a non-contact setting by using cameras (Takano & Ohta, 2007; Verkruysse, Svaasand & Nelson, 2008). This technique estimates physiological signals by analyzing subtle, periodic changes in skin reflectance caused by blood volume fluctuations. It is based on the principles of contact-based photoplethysmography (PPG), which captures blood volume pulse (BVP) through optical reflectance or absorbance, and instead leverages standard RGB video to extract pulsatile signals. As hemoglobin concentration in the superficial microvasculature varies during each cardiac cycle, it alters the absorption and reflection of ambient light, producing imperceptible color changes on the skin surface. rPPG captures these changes by measuring pixel intensity variations across frames, primarily in facial or thoracoabdominal regions, and reconstructs waveforms that correlate with heart rate and respiratory motion. Effective implementation requires stable anatomical regions with sufficient skin exposure, adequate temporal resolution to capture physiological frequencies, and robust signal conditioning steps such as detrending, band-pass filtering, and color-space projections to enhance pulsatility.

1.1.3 Clinical Decision-Making Framework in PICUs

In pediatric intensive care units (PICUs), vital sign monitoring plays a central role in guiding clinical decision-making across multiple time scales. These physiological measurements inform early detection of deterioration, support titration of therapeutic interventions, and provide continuous assessment of treatment response. Scoring systems such as the Pediatric Early Warning Score (PEWS) integrate parameters like heart rate and respiratory rate to predict clinical decline. Prior studies have demonstrated that abnormal PEWS values frequently precede critical events, including cardiac arrest, by several hours (Parshuram & et al., 2011). This anticipatory window underscores the clinical imperative for uninterrupted, high-fidelity monitoring that can detect subtle physiological shifts in real time.

A major challenge in translating advanced monitoring systems into the PICU context is the pervasive issue of alarm fatigue. Clinical audits report that each patient may trigger between 150 and 400 alarms daily, with false positive rates often exceeding 85 percent for key parameters (Graham & Cvach, 2010; Cvach, 2012). This excessive alarm burden leads to desensitization among clinical staff, potentially delaying intervention during genuine emergencies.

Therefore, monitoring systems must not only achieve high sensitivity in detecting vital sign fluctuations, but also demonstrate sufficient specificity to suppress spurious alerts. Temporal stability becomes equally critical; systems that produce jittery or inconsistent outputs may introduce more clinical noise than benefit. As such, signal robustness and alarm reliability must be considered core design criteria for any monitoring approach intended for clinical deployment in the PICU.

1.2 Methodological Paradigms in Medical Video Analysis

Two main approaches exist for estimating the rPPG signal: traditional signal processing techniques and deep learning-based models. Early methods rely on detecting the face and extracting regions of interest (ROIs) that exhibit color variations caused by pulsatile blood flow at the skin surface. These methods process temporal signals obtained from the red, green, and blue (RGB) channels of the video to derive an estimate of the blood volume pulse. In contrast, deep learning-based approaches decompose the task into spatial and temporal components. The spatial module, often implemented using convolutional neural networks (CNNs), captures relevant features from facial regions such as the forehead and cheeks, which are known to exhibit prominent photoplethysmographic signals. The temporal module models frame-to-frame dependencies to track physiological dynamics over time, enabling robust signal reconstruction under variable conditions.

1.2.1 Traditional Signal Processing Techniques

Traditional methods for rPPG estimation typically follow a two-stage pipeline. The first stage involves detecting the face and identifying skin regions within the facial area that are suitable for signal extraction. In the second stage, pixel intensities are collected from these regions, and signal processing techniques are applied to extract the rPPG signal. This involves analyzing temporal fluctuations in the red, green, and blue (RGB) channels to isolate periodic components corresponding to the blood volume pulse (BVP).

Conventional signal processing methods for rPPG estimation operate without learning-based supervision and follow a well-defined pipeline: detection of facial or skin regions, spatial averaging over selected pixels, temporal filtering to suppress non-pulsatile fluctuations, and signal projection to enhance the physiological component. Techniques such as Independent Component Analysis (ICA) (Poh, McDuff & Picard, 2010a) and POS(Wang, den Brinker, Stuijk & de Haan, 2017b), aim to separate the cardiac pulse from other overlapping sources by decomposing the RGB channels into statistically independent signals. Chrominance-based approaches, including CHROM (De Haan & Jeanne, 2013), mitigate sensitivity to lighting by transforming the signal into a color subspace less affected by illumination changes. Similarly, the Plane-Orthogonal-to-Skin (POS) method projects the RGB signal onto a plane orthogonal to a normalized skin tone, attenuating motion and lighting artifacts (Wang, den Brinker, Stuijk & de Haan, 2017a). These methods are computationally efficient and interpretable, making them suitable baselines in low-resource settings.

Despite their effectiveness in controlled environments, traditional handcrafted rPPG pipelines rely on strict assumptions about lighting conditions, skin reflectance properties, and stable subject positioning. These constraints limit their applicability in clinical contexts such as the Pediatric Intensive Care Unit (PICU), where non-stationary illumination, occlusions, motion artifacts, and sensor noise are frequent and often unavoidable. In practice, these factors introduce non-physiological variance that degrades signal quality and propagates through each stage of the processing pipeline, ultimately compromising the reliability of the estimated waveform. Moreover, region-of-interest (ROI) instability caused by patient movement or pose variation further exacerbates signal distortion, particularly in neonates and infants with irregular morphology. As a result, these methods struggle to generalize across patients and clinical conditions, motivating a shift toward learning-based models that can better accommodate the variability inherent in real-world pediatric care environments.

1.2.2 Deep Learning Approaches for rPPG Estimation

To overcome the limitations of handcrafted pipelines, recent approaches adopt end-to-end deep learning models that estimate rPPG signals directly from raw video inputs. Convolutional networks capture local spatial patterns, while Transformers and state space models enable efficient modeling of long-range temporal dynamics, eliminating the need for separate signal-processing stages.

1.2.2.1 CNN-Based Models

Early deep learning approaches in rPPG relied on Convolutional Neural Networks (CNNs) for feature extraction from facial videos. Initial approaches applied 2D CNNs on individual video frames to learn spatial features. For example, DeepPhys (Chen & McDuff, 2018) introduced a motion representation guided by an attention mechanism to estimate heart rate (HR) using skin reflection models. Other methods combined video magnification with deep learning, as in Eulerian Video Magnification (EVM)-based approaches, which filtered spatial and temporal signals before feeding them into CNNs for HR estimation (Qiu, Liu, Arteaga-Falconi, Dong & El Saddik, 2019a). The HR-CNN model integrated a two-part structure with a 2D CNN for spatial feature extraction and an estimator module for HR prediction. However, such designs were limited by their inability to model temporal relationships across frames (Špetlík, Franc & Matas, 2018). EfficientPhys addressed some of these issues by discarding facial landmark preprocessing and incorporating a normalization module with 2D convolution layers, allowing efficient spatiotemporal feature extraction directly from raw video (Liu, Hill, Jiang, Patel & McDuff, 2022a). MTTS-CAN proposed the use of a Temporal Shift Module (TSM) to enhance temporal signal encoding by averaging adjacent frames before feeding them into the network, improving temporal coherence while suppressing motion noise (Liu, Fromm, Patel & McDuff, 2020). These early 2D-CNN approaches demonstrated the feasibility of camera-based pulse detection, but they processed video frames independently and struggled with temporal context and motion noise. To address these issues, PulseGAN proposed a generative adversarial model that takes a rough rPPG signal and denoises it to produce a clean ppg waveform, significantly improving heart rate estimation under challenging conditions (Song *et al.*, 2021).

Among the limitations of basic 2D CNN models was their inability to leverage information across time. The shift toward 3D CNNs brought significant improvements in rPPG signal estimation by learning spatial and temporal features simultaneously. rPPGNet (Yu, Peng, Li, Hong & Zhao, 2019c) was an early 3D-CNN model that included a video enhancement to handle compression artifacts, thus feeding higher-quality inputs into the rPPG estimator. PhysNet employed a fully 3D-CNN architecture to generate spatiotemporal feature representations tailored to rPPG (Yu, Li & Zhao, 2019a). AutoHR (Yu, Li, Niu, Shi & Zhao, 2020) employs neural architecture search to optimize a 3D CNN with temporal difference convolutions, incorporating hybrid loss and data augmentation for accurate HR measurement across datasets. Other innovations augmented 3D CNNs with attention modules to focus on pertinent signals. DeeprPPG proposed a lightweight CNN model employing 3D spatiotemporal convolutions to aggregate pulse information over consecutive frames (Liu & Yuen, 2020). Other approaches augmented 3D CNNs with attention modules to focus on pertinent signals. ETA-rPPGNet (Hu *et al.*, 2021a) introduced a time-domain segmentation module and attention mechanisms to isolate temporal cues. Similarly, SAM-rPPGNet (Hu *et al.*, 2021b) leveraged a spatial-temporal attention design, including strip pooling and a spatial attention module (SAM), to reduce the impact of head motion and facial dynamics. As efficiency became a concern, some proposed techniques optimize 3D CNNs for real-time use. RTrPPG (Botina-Monsalve, Benezeth & Miteran, 2022) exemplified this by finding a sweet spot between accuracy and speed: it reduced input resolution and model size and introduced a hybrid loss function that combines time- and frequency-domain objectives with an SNR penalty

Hybrid CNN-RNN models were proposed to further capture temporal dependencies. RhythmNet (Niu, Shan, Han & Chen, 2020a) used a 2D CNN to learn spatial features, followed by a Gated Recurrent Unit (GRU) to model temporal sequences. Although effective in capturing short-term dependencies, these architectures introduced substantial computational complexity and often failed to generalize long-term temporal relationships. Another hybrid design employed

ConvLSTM layers after CNN-based feature extraction, enabling end-to-end training for rPPG signal reconstruction (Hu, Guo, Wang, Ge & Chu, 2019). Meta-rPPG (Lee, Chen & Lee, 2020b) extended this by introducing a metalearner for domain adaptation, combining a 2D CNN with a BiLSTM backbone to generalize across subjects and recording conditions (Lee, Chen & Lee, 2020a). EVM-CNN (Qiu, Liu, Arteaga-Falconi, Dong & Saddik, 2019b) estimates heart rate in real time by amplifying subtle facial color variations with Eulerian Video Magnification and predicting the pulse using a lightweight CNN. HeartTrack (Perepelkina, Artemyev, Churikova & Grinenko, 2020) uses a convolutional neural network to extract subtle spatiotemporal facial cues from video and estimate heart rate remotely and in real time. ETA-rPPGNet (Hu *et al.*, 2021c) estimates heart rate from facial video by applying time-domain attention to highlight informative temporal variations in the rPPG signal. PRNet (Huang, Lin, Chen, Juang & Wu, 2021a) adopted a unified 3D-CNN and LSTM framework, enabling both spatial and temporal modeling in a single-stage pipeline. Yet, comparative evaluations confirmed that purely 3D CNN models often outperform CNN-RNN hybrids by eliminating sequential modeling overhead while preserving signal fidelity (Yu *et al.*, 2019a). The sequential modeling in RNNs adds overhead and risk of overfitting, whereas a well-trained 3D CNN can capture the necessary temporal features in its convolutional filters.

Despite recent progress, supervised deep learning approaches for rPPG estimation remain limited by three key challenges that hinder clinical deployment. First, they are heavily dependent on large-scale annotated datasets, which are rarely available in clinical settings due to privacy constraints and high labeling costs. While public datasets like VIPL-HR include thousands of samples, they are often captured in controlled environments with cooperative subjects and fail to reflect the complexity of real-world data. Second, conventional CNNs are constrained by local receptive fields, which limits their ability to model global spatial dependencies. In scenarios with frequent occlusions, such as when only parts of the face are visible, these networks struggle to integrate dispersed pulse information across larger regions. Third, temporal modeling is typically restricted to short sequences, with 3D convolutions capturing only local dynamics over 3–5 frames. Given that physiological signals span longer durations, these methods lack the capacity to represent long-term temporal patterns without significantly increasing model depth and computational cost.

1.2.2.2 Transformer-Based Models

Vision Transformers (ViTs) (Dosovitskiy *et al.*, 2020) address the local-receptive-field limitation of CNNs through multi-head self-attention, which computes pairwise interactions between all tokens in a sequence. This enables global context modeling: every spatial patch can attend to every other patch, and in video transformers every frame can interact with every other frame. For rPPG, this could theoretically allow the model to integrate pulse information across the entire face and across long temporal windows.

Several works have explored transformers for rPPG. PhysFormer (Yu *et al.*, 2022) proposed a temporal difference attention mechanism that emphasizes subtle inter-frame changes while leveraging global spatiotemporal attention. The model was further extended in PhysFormer++, which employed a SlowFast dual-branch Transformer to capture both fast-varying and slow global trends (Yu *et al.*, 2023). TransPPG (Kang, Yang & Zhang, 2024) introduced a two-stream architecture that processes facial regions and background separately, enabling the model to suppress noise caused by lighting and motion artifacts. EfficientPhys (Liu *et al.*, 2022a) adapted the Swin Transformer with Temporal Shift Modules (TSMs) to enable lightweight attention over spatial patches while efficiently propagating temporal information across frames. Revanur et al. (Revanur, Dasari, Tucker & Jeni, 2022) leverages video transformers to estimate instantaneous physiological measurements, including heart rate. The model captures spatial and temporal dynamics using self-attention mechanisms, enabling the extraction of fine-grained temporal features directly from video sequences.

TimeSformer (Bertasius, Wang & Torresani, 2021) introduced a divided space-time attention mechanism that separately models spatial and temporal dependencies. rPPGTR (Zhang, Yang, Yin & Meng, 2023a) integrated local CNN-based features with TimeSformer attention blocks to enhance signal extraction under variable conditions. RhythmFormer (Zou *et al.*, 2025) introduced periodic sparse attention to reduce the computational burden of full self-attention, filtering out irrelevant temporal frames. Spiking-PhysFormer (Liu *et al.*, 2025) replaced conventional self-attention with spike-driven neuromorphic modules to reduce power consumption for mobile applications.

Although transformer-based architectures have shown strong performance on benchmark datasets, several limitations restrict their applicability in clinical environments. A primary challenge is their quadratic complexity, as standard self-attention scales with the square of the sequence length. For video-based physiological monitoring, where clips span hundreds of frames and cover high spatial resolutions, the resulting token count renders full attention computation prohibitively expensive. Even with optimizations such as divided space-time attention, the memory and latency overhead remain incompatible with real-time bedside requirements. In addition, transformers exhibit weaker inductive biases than convolutional networks. Lacking built-in spatial priors such as translational equivariance, they require extensive pretraining to learn representations that CNNs capture more naturally. This dependency on large labeled datasets poses a critical barrier in medical contexts, where data is limited and costly to annotate. Furthermore, transformers struggle under occlusion. While attention mechanisms can in theory downweight irrelevant regions, standard implementations lack explicit mechanisms to identify and prioritize informative visible areas. When facial visibility is reduced attention is often distributed inefficiently, diminishing robustness and limiting clinical utility.

1.2.2.3 State-space models (SSMs)

State-space models (SSMs) have emerged as a compelling alternative to transformer architectures in sequence modeling tasks, offering linear-time complexity while effectively capturing long-range dependencies. Unlike transformers, which rely on pairwise attention computations that scale quadratically with sequence length, SSMs parameterize hidden states through linear recurrence relations, enabling efficient processing of extended temporal contexts. The foundational Mamba architecture (Gu & Dao, 2023) introduced a selective mechanism that conditions state transitions on input data, bridging the gap between recurrent and attention-based models. This selective SSM formulation has demonstrated competitive performance across language and vision tasks.

In the domain of video understanding and rPPG estimation, SSMs address the computational bottlenecks of transformers by decoupling spatial and temporal processing. VideoMamba (Li *et al.*, 2024) adapted SSMs for efficient video feature extraction, achieving real-time inference on long clips while preserving global context. For rPPG specifically, RhythmMamba (Zou, Guo, Hu & Ma, 2024) leveraged SSMs to model periodic physiological signals, capturing inter-frame pulsatile dynamics with linear complexity and outperforming CNN baselines on benchmarks like UBFC-rPPG. PhysMamba (Luo, Xie & Yu, 2024) introduced a dual-path SSM framework with temporal difference blocks, fusing time and frequency domains for robust HR estimation under motion artifacts. Similarly, CardiacMamba (Wu *et al.*, 2025) integrated SSMs in a multimodal RGB-RF fusion pipeline, enhancing signal denoising and achieving low MAE in cross-modality scenarios. ME-rPPG (Zhang, Lu, Liu, Chen & Wu, 2025) focused on memory efficiency through temporal-spatial SSM duality, reducing latency for mobile applications while maintaining clinical-grade accuracy.

These SSM-based approaches offer distinct advantages for rPPG in resource-constrained environments like PICUs, including lower memory footprints and faster inference times compared to transformers. While effective in controlled datasets, their generalization to diverse clinical conditions remains underexplored, particularly in pediatric populations with variable skin tones and morphologies. Despite these advancements in deep learning architectures for rPPG, persistent challenges in data scarcity and domain adaptation necessitate complementary strategies such as self-supervised learning, which can leverage unlabeled clinical videos to enhance model robustness.

1.3 Theoretical Foundations

This thesis relies on three core mathematical frameworks to address the challenges of physiological signal extraction, efficient tracking, and self-supervised learning. To clarify the methodological choices presented in Chapters 2, 3, and 4, we detail here the physical and mathematical intuition behind the key equations governing our models.

1.3.1 Modeling Pulsatile Dynamics: Temporal Central Difference Convolution

Standard 3D convolutions aggregate features based on absolute intensity, rendering them sensitive to static skin tone and ambient illumination. In rPPG, the physiological signal of interest is embedded in intensity *variations* rather than the intensity itself. To capture these fine-grained dynamics, we employ Temporal Central Difference Convolution (3DCDC-T), which integrates an explicit gradient term into the aggregation process:

$$y(p_0) = \theta \cdot \sum_{p_n \in \mathcal{R}} w(p_n) \cdot \big(x(p_0 + p_n) - x(p_0)\big) + (1 - \theta) \cdot \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n) \tag{1.1}$$

where p_0 denotes the current spatiotemporal location and $\mathcal{R}$ represents the local receptive field. The function $x(\cdot)$ corresponds to the input feature map, $w(p_n)$ denotes the learnable convolutional weight associated with the relative position p_n, and $\theta \in [0, 1]$ controls the balance between the gradient-based term and the standard intensity-based aggregation. The difference term $(x(p_0 + p_n) - x(p_0))$ encodes local spatiotemporal variations around the center p_0. Incorporating this differential component makes the operator more sensitive to the subtle reflectance changes induced by the cardiac cycle.

1.3.2 Global Context Capture: Self-Attention Mechanism

To capture long-range dependencies across the video sequence, we initially consider the standard Multi-Head Self-Attention (MSA) mechanism found in Vision Transformers. Unlike convolutions, which are limited to local receptive fields, MSA allows the model to relate physiological features at any time step t to all other time steps in the sequence.

Given an input sequence $X \in \mathbb{R}^{L \times D}$, the transformer projects it into queries (Q), keys (K), and values (V). The attention output is computed as a weighted sum of values, where weights are determined by the compatibility between queries and keys:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1.2}$$

where Q (queries), K (keys), and V (values) are the projected input matrices, and d_k is the key dimensionality used as a scaling factor in the softmax to stabilize the attention scores. While this mechanism effectively captures global physiological trends, it requires computing a similarity matrix of size $L \times L$, leading to a quadratic complexity of $O(L^2)$.

1.3.3 Efficient Sequence Modeling: Mamba Discretization

In Chapter 3, the model shifts from transformer-based attention to state space formulations to achieve linear-time sequence processing for video detection. The Mamba architecture relies on a discretization of the continuous-time system $h'(t) = Ah(t) + Bx(t)$ into a form suitable for digital sequences. This is realized through a learnable timescale parameter Δ, applying the Zero-Order Hold (ZOH) principle:

$$\bar{A} = \exp(\Delta A), \quad \bar{B} = (\Delta A)^{-1}\big(\exp(\Delta A) - I\big) \cdot \Delta B \tag{1.3}$$

where $\overline{A}$ and $\overline{B}$ are the discretized state parameters, A and B denote the continuous-time system matrices, and I is the identity matrix. In physiological video analysis, the parameter Δ acts as a content-dependent temporal resolution: it determines how information is integrated across time, effectively controlling which historical features are preserved or forgotten. This adaptive discretization enables the model to capture temporal structure with the fidelity of recurrent formulations while retaining the parallel training efficiency characteristic of convolutional and feed-forward architectures.

1.4 Domain Adaptation and Transfer Learning in Clinical Settings

1.4.1 The Lab-to-Clinic Gap in Non-Contact Monitoring

Remote photoplethysmography (rPPG) has demonstrated promising results in laboratory environments, where conditions such as controlled lighting, minimal motion, and cooperative subjects facilitate robust signal extraction. Models trained on benchmark datasets such as UBFC-rPPG or PURE typically perform well in these ideal settings, as they are optimized for clean, periodic signals. However, performance deteriorates when these models are deployed in clinical environments, particularly in pediatric intensive care units (PICUs), where real-world constraints introduce substantial noise, artifacts, and inter-patient variability. This discrepancy, widely recognized as the lab-to-clinic gap, highlights the structural differences between controlled experiments and clinical workflows, posing a significant barrier to clinical translation (Huang *et al.*, 2023; Nagar, Hasegawa-Johnson, Beiser & Ahuja, 2024).

In the PICU, the visual complexity of recorded scenes arises from multiple, interrelated factors. Patients frequently exhibit involuntary movements or are repositioned by medical staff, and they present a wide range of facial anatomies and skin tones. Occlusions caused by oxygen masks, ventilator tubes, and bedding are common and systematic rather than incidental (Svoboda, Sperrhake, Nisser, Taphorn & Proquitté, 2024). PICU dataset analysis indicates that medical equipment obscures, on average, 23% of the facial region, with the nose and mouth, which are key landmarks for respiratory signal extraction, being the most frequently covered. In addition, lighting conditions in the PICU are non-uniform and may shift abruptly due to bedside procedures, while camera angles are often suboptimal or misaligned. These factors compound the domain mismatch and result in degraded model generalization when moving from laboratory data to clinical deployment (Finlayson *et al.*, 2021).

Demographic and physiological mismatches between training and deployment domains further exacerbate this performance gap (Bondarenko, Menon & Elgendi, 2025). Public rPPG datasets largely consist of healthy adults captured in static, frontal poses under uniform illumination. In contrast, PICU patients span a diverse range of ages, clinical conditions, and skin types. Many are sedated or critically ill, and their facial features may be underdeveloped or altered by disease. These disparities lead models to overfit to lab-specific distributions, resulting in substantial loss of accuracy when applied to heterogeneous clinical data (Xie, Yu, Wu, Xie & Shen, 2024; Sun *et al.*, 2023). The consequences include higher heart rate estimation errors and diminished signal quality, both of which undermine the reliability and clinical utility of non-contact monitoring in real-world pediatric care settings.

1.4.2 Data Scarcity, Ethical Constraints, and Domain Generalization

Developing clinically robust models for non-contact vital sign monitoring in pediatric intensive care units (PICUs) is fundamentally constrained by limited access to representative data, stringent ethical regulations, and the challenge of generalizing across heterogeneous clinical environments. In pediatric research, the collection of large-scale annotated video datasets is inherently difficult due to both logistical barriers and ethical considerations. Informed consent protocols typically involve guardians who must evaluate the potential research benefits in light of concerns over privacy, safety, and emotional impact (Muralidharan, Burgart, Daneshjou & Rose, 2023). These constraints reduce not only the volume of data available but also its representativeness, as approvals are more likely for patients with less severe conditions, introducing selection bias into the dataset.

Existing public datasets such as UBFC-rPPG and PURE, while widely used for benchmarking, do not reflect the complexity of the clinical setting (Nguyen, Nguyen, Li, López & Casado, 2024). These datasets are collected in controlled laboratory environments with adult participants, uniform lighting conditions, and minimal movement. In contrast, the PICU environment is marked by dynamic lighting changes, frequent occlusions from medical equipment, and spontaneous patient movements (Ruhrberg Estévez *et al.*, 2025). Currently, no publicly available dataset captures both the facial and thoracoabdominal regions under these realistic conditions. As a result, models trained solely on laboratory data tend to generalize poorly, with rPPG estimation errors often exceeding clinically acceptable thresholds. Supervised models, which rely heavily on annotated clean signals, are particularly vulnerable when exposed to pediatric-specific challenges such as immature vascular structures, diverse skin tones, and irregular respiratory rhythms (Huang *et al.*, 2023).

Ethical principles further restrict data acquisition. In pediatric care, the principle of non-maleficence discourages prolonged or intrusive monitoring unless it offers direct clinical benefit (Snyder, Stewart & Hunter, 2025). This ethical stance limits opportunities to record extended video sequences for research purposes. When such recordings are approved, the requirements for anonymization must strike a balance between preserving physiological signal fidelity and protecting patient identity. This often results in fragmented or inconsistently preprocessed datasets (Abusamra *et al.*, 2025; Nisevic, Milojevic & Spajic, 2025). Furthermore, manual annotation is resource-intensive, making large-scale supervised learning impractical in most pediatric hospital settings.

To address these limitations, recent work has increasingly focused on self-supervised learning strategies that reduce reliance on labeled data. Techniques such as masked autoencoding enable models to learn spatiotemporal representations directly from unlabeled video, improving robustness to occlusion, motion artifacts, and signal noise (Liu *et al.*, 2024b). Contrast-Phys (Sun & Li, 2022) learns remote physiological signals from video in an unsupervised way by using spatiotemporal contrast to distinguish pulse-related color changes from background variations. In parallel, domain generalization is supported through curriculum-based training strategies that gradually expose the model to increasing complexity (Wang, Islam, Xu & Ren, 2024a; Singh, Nampalle, Narayan & Raman, 2023). Rather than depending on domain adaptation techniques that assume access to target-domain data, these curricula begin with clean laboratory data, introduce synthetic occlusions and distortions, and finally incorporate unlabeled clinical recordings for pretraining. This approach aligns with ethical constraints while also improving the generalizability of models deployed in the high-variance and visually complex environment of the PICU (Sun *et al.*, 2023).

1.5 Clinical Deployment Considerations

1.5.1 Real-Time Processing Requirements in Clinical Settings

Integrating remote photoplethysmography (rPPG) into pediatric intensive care units (PICUs) requires alignment between computational efficiency and compatibility with clinical workflows. Unlike controlled experimental setups, clinical environments impose strict requirements on latency, reliability, and resource utilization. In this context, real-time processing is essential, not optional, as the timely detection of physiological changes supports rapid clinical decision-making. Scoring systems such as the Pediatric Early Warning Score (PEWS) depend on continuous updates of heart rate and respiratory rate to evaluate patient stability (Parshuram & et al., 2011).

For clinical deployment, systems must meet deterministic latency requirements, where any variability in processing time may disrupt synchronization with other monitoring tools or delay alarm triggering. Signal estimation must be initialized within 8 to 10 seconds of recording onset, followed by updates at a 1 Hz rate to remain aligned with clinical response windows (Villarroel *et al.*, 2019). Models that achieve high average throughput but occasionally exhibit latency spikes fall short of this requirement, as such unpredictability introduces operational risks.

In addition, the hospital computing environment presents additional constraints. Clinical workstations often run multiple resource-intensive applications simultaneously, including electronic health records, imaging software, and alert dashboards. Any rPPG solution must operate alongside these systems without degrading performance. This necessitates architectures that run efficiently on standard CPUs or embedded platforms, which typically lack high-end GPUs. Optimizing memory usage and compute efficiency is essential to avoid bottlenecks or delayed feedback (Álvarez Casado, Padilla-López, Latorre-Crespo, Pérez-Borràs & Casas, 2023b).

Several implementations have addressed these challenges through system-level and architectural optimizations. Multithreaded pipelines are commonly used to parallelize video acquisition, signal estimation, and data transmission. Lightweight 3D convolutional neural networks (CNNs) and deeply supervised models have achieved throughput between 30 and 60 frames per second on mid-range devices. These results were enabled by strategies such as input resolution downsampling, hybrid time-frequency loss functions, and pruning (Botina-Monsalve *et al.*, 2022; Lee, Lee & Sim, 2023a). More recently, energy-efficient prototypes have introduced early feature truncation and memoization to support real-time operation on portable or bedside hardware.

From a model design perspective, transformer-based architectures impose substantial computational overhead due to their quadratic complexity and large parameter sizes. This limits their practicality for long video sequences in resource-constrained environments. In contrast, state-space models offer a scalable alternative, providing linear time complexity while retaining the capacity for long-range temporal modeling (Zou *et al.*, 2024). Their lower memory and computational footprint makes them especially suitable for hospital settings with tight resource budgets and low latency tolerance.

In summary, deploying rPPG in the PICU requires addressing a convergence of challenges, including strict latency demands, limited computational resources, and integration with complex clinical workflows. The methods proposed in this thesis, including linear-time state-space architectures, adaptive video masking, and curriculum-based training, are designed with these constraints in mind. By tackling both the algorithmic and deployment-level challenges, the proposed framework aims to enable robust and clinically viable real-time non-contact monitoring.

1.5.2 Regulatory Frameworks and Validation Requirements

The deployment of AI-based monitoring systems in clinical environments must comply with stringent regulatory frameworks that directly influence system design, evaluation protocols, and clinical integration. In the United States, the Food and Drug Administration (FDA) regulates AI-enabled medical devices, including software as a medical device (SaMD) intended for vital sign monitoring. Under the FDA's risk-based classification system, most AI-based monitoring tools fall under Class II or III, necessitating either a 510(k) premarket notification or a premarket approval (PMA) process (U.S. Food and Drug Administration, 2019). For adaptive models that evolve post-deployment, the FDA's 2021 Action Plan introduced the concept of predetermined change control plans (PCCPs), which enable manufacturers to update AI systems while maintaining compliance without repeated regulatory submissions (U.S. Food and Drug Administration, 2021). More recent draft guidance further elaborates on lifecycle management, recommending that developers address issues such as algorithmic transparency, bias mitigation, and validation across diverse patient populations (U.S. Food and Drug Administration, 2025).

These evolving requirements pose specific challenges for video-based rPPG systems. Establishing substantial equivalence is difficult when the core sensing modality departs from established contact-based techniques. In such cases, developers must present novel evidence frameworks that both demonstrate safety and efficacy and align with regulatory expectations for innovative technologies. For pediatric applications in particular, the FDA emphasizes the inclusion of age-stratified validation data, acknowledging developmental variability in skin reflectance, heart rate ranges, and respiratory patterns that can affect measurement reliability. Consequently, regulatory constraints guide technical decisions throughout the model development pipeline. Validation must not only report global accuracy metrics but also establish robustness across clinically relevant subgroups. In the case of heart rate monitoring, performance is typically benchmarked using ANSI/AAMI EC13 standards, which permit an absolute error of ±10% or ±5 beats per minute (bpm), whichever is greater, across a 30–250 bpm range (Association for the Advancement of Medical Instrumentation, 2002). These thresholds are particularly relevant in pediatric populations, where baseline heart rates are higher and rapid fluctuations are common.

Recent clinical studies have begun to address these regulatory criteria, providing early evidence for rPPG's feasibility in pediatric care. A cross-sectional study of neonates and children up to 16 years demonstrated high correlation in heart rate estimation, particularly among older children (Spearman's rank correlation coefficient of 0.82) (Hatib *et al.*, 2024). However, the study highlighted the need for refinement in respiratory rate and oxygen saturation estimation, especially among neonates. Complementary evidence from a systematic review and meta-analysis reported low mean biases in contactless rPPG-derived heart rate (-0.25 bpm) and respiratory rate (0.65

bpm), reinforcing its potential to reduce alarm fatigue and skin injury risks (Bautista *et al.*, 2023). Nonetheless, the review emphasized the lack of standardized reporting procedures, which complicates regulatory evaluation and clinical translation.

Despite incremental progress, significant challenges remain. Publicly available clinical datasets remain scarce, and many existing models lack built-in mechanisms to quantify prediction confidence, detect data drift, or handle cases with poor visibility due to occlusions or motion artifacts. FDA guidance and clinical safety considerations increasingly favor AI systems that offer real-time reliability assessments, such as logging of uncertainty estimates and alerting clinicians in cases where model output may be compromised. Attention-based visualization tools provide intuitive interpretability, offering clinicians insight into why a given prediction was made (Benjamens, Dhunnoo & Meskó, 2020).

Future validation efforts should prioritize the development and evaluation of rPPG systems on datasets reflecting diverse skin tones, clinical conditions, and care environments to mitigate bias and support equitable deployment. By aligning with established regulatory standards and integrating clinician-centered evaluation frameworks, video-based rPPG can progress toward safe and effective adoption in PICUs.

1.5.3 Human-Computer Interaction in Critical Care

In pediatric intensive care units (PICUs), the interface between artificial intelligence (AI) systems and clinical users fundamentally shapes practical utility, regardless of the system's underlying technical accuracy. Monitoring platforms often serve as the primary conduit through which clinicians access and interpret physiological data, yet suboptimal interface design can contribute to cognitive overload, misinterpretation, and delays in intervention (Patel & Buchman, 2016; Hulyalkar *et al.*, 2017). Beyond presenting accurate vital signs, these systems must convey information that is contextualized and immediately actionable, aligning with established clinical workflows and decision-making heuristics. In particular, the visualization of video-derived vital signs should respect clinicians' mental models; outputs that diverge from expected patterns without adequate explanation may undermine trust, even when technically correct.

Alarm fatigue remains one of the most pervasive human-computer interaction challenges in critical care. Studies estimate that up to 94% of alarms in intensive care contexts are non-actionable, leading clinicians to routinely ignore or silence them and increasing the risk of missed critical events (Bonafide *et al.*, 2015; Huo, Wung, Roveda & Li, 2023). Integrating video-based monitoring may inadvertently amplify this burden, triggering additional alerts related to signal quality degradation, detection failures, or physiological abnormalities. To mitigate this, user-centered design approaches recommend adaptive alarm mechanisms that incorporate clinical context, including patient-specific parameters such as age, baseline trends, and current treatment regimens (van Rossum *et al.*, 2022). Emerging visualization strategies that consolidate multiple vital signs into unified timelines with color-coded deviations have demonstrated reduced time to detect deterioration by 30% in adult ICUs (Görges, Kück, Koch, Agutter & Westenskow, 2011), supporting the potential value of similar approaches tailored for pediatric use.

Effective trust calibration, defined as the clinician's appropriate reliance on AI-assisted outputs, requires transparent communication of uncertainty and system limitations. Surveys of PICU providers have highlighted concerns regarding opacity in algorithmic decision-making, with many expressing hesitation toward adopting AI tools without interpretable justifications for predictions (Al-Sofyani, 2025; Popoff, Cabon, Cuggia, Bouzillé & Clavier, 2025). In this context, systems must not only display measurement values but also indicate their reliability under varying conditions. For example, our approach separates face and thorax detection pipelines, enabling the system to provide independent confidence scores for each anatomical region. When facial occlusion precludes accurate cardiac signal extraction, thoracic cues can still support respiratory rate estimation, allowing the system to preserve partial functionality. The incorporation of explainable AI components, such as saliency heatmaps that highlight spatial contributions to predicted vital signs, further enhances user understanding and confidence (Rohmetra *et al.*, 2023; Hossain, Muhammad & Guizani, 2020).

Ultimately, improving human-computer interaction in the PICU context demands iterative development cycles that involve clinical end-users from the outset. Usability testing in operational environments, rather than simulated settings, remains critical for ensuring that novel monitoring solutions translate to tangible improvements in workflow efficiency and patient safety (Helman *et al.*, 2022).

CHAPTER 2

RPPG ESTIMATION : VISION TRANSFORMER WITH 3D TEMPORAL CENTRAL DIFFERENCE

Mohamed Khalil Ben Salah[1] , Philippe Jouvet[2] , Rita Noumeir[1]

[1] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[2] Research Center at CHU Sainte-Justine Hospital, University of Montreal,
3175 Chem. de la Côte-Sainte-Catherine, Montréal, Québec, Canada H3T 1C5

Article published in « IEEE Transactions on Instrumentation and Measurement » in November 2024.

2.1 Abstract

Remote photoplethysmography (rPPG) has gained increasing importance, especially during and after the COVID-19 pandemic, for its ability to estimate heart rate by analyzing subtle variations in skin color without physical contact. This non-invasive and practical method relies on capturing changes in pixel intensity through RGB video or near-infrared imaging. In this study, we propose a novel hybrid model that leverages the feed-forward integration of 3D Convolutional Neural Networks (3D-CNNs) and Video Vision Transformers (ViViTs) to enhance rPPG estimation. The 3D-CNNs first capture local spatiotemporal features, using Temporal Central Difference Convolution (3DCDC-T) and Convolutional Block Attention Module (CBAM). These local features are then passed to the Video Vision Transformer, where Multi-Head Self-Attention (MHSA) captures global contextual relationships and long-range dependencies across frames, enabling more effective representation of complex temporal dynamics. This sequential learning allows the model to progressively refine features from local to global, ensuring more consistent and coherent feature extraction. Our feed-forward approach also improves computational efficiency by reducing the dimensionality of the input data before global attention processing, making it particularly effective in data-limited environments. Through comprehensive experiments, we show that our hybrid approach outperforms state-of-the-art methods across multiple public datasets, achieving a 22.55% improvement in MAE and a 55.80% improvement in RMSE on the

UBFC-rPPG dataset, demonstrating superior feature progression and generalization in rPPG and heart rate estimation tasks.

2.2 Introduction

Currently used methods for measuring heart rate, such as electrocardiograms (ECGs), often demand physical contact with a patient and could be costly and complex to wear; these factors can limit their availability and widespread adoption. However, recent advancements have demonstrated the feasibility of measuring heart rate remotely through a technique known as remote photoplethysmography (rPPG). The technique of rPPG utilizes ubiquitous sensors, such as cameras, to analyze images and videos and capture subtle changes in the light reflected by the skin. The variations in blood volume caused by cardiac rhythm allow us to extract vital signals. This technique provides a non-contact method for measuring vital signs, primarily heartbeats, in situations where direct physical contact with the person is challenging, not recommended, or prohibited. The application of rPPG is particularly beneficial in hospital settings, where it can be used for monitoring normal patients as well as those with specific conditions such as sensitive children, newborns, individuals with skin lesions, or patients with highly contagious diseases (for example COVID-19). By avoiding direct contact, rPPG offers a non-invasive and convenient means of monitoring vital signs in various scenarios where traditional methods may not be feasible or appropriate.

There are two approaches for estimating rPPG signals: first, there are traditional methods based on signal processing, and then there are deep learning approaches, which have shown promising results. In the earliest approaches, rPPG signal is estimated after applying face detection techniques, then extracting the region of interest to get color variations induced by the impulses on the surface of the human skin. By processing the received signals from the red, green, and blue channels, they employed signal processing methods to derive an estimated signal of remote photoplethysmography (rPPG); Blind source separation (BSS) (Poh, McDuff & Picard, 2010b) is one of these methods which is based on decomposition of the RGB signal, and chrominance-based method (De Haan & Jeanne, 2013) converts the RGB color space into the chrominance domain. Moreover, plane-orthogonal-to-skin (POS) (Wang *et al.*, 2017b) projects the RGB signals onto a plane orthogonal to a normalized skin tone in the normalized RGB space.

The deep learning approaches include two modules: spatial and temporal. Convolutional Neural Networks (CNNs) enable the model to obtain spatial information after extracting regions of the face containing relevant information about vital signs (Chen & McDuff, 2018; Yu, Peng, Li, Hong & Zhao, 2019d; Yu, Li & Zhao, 2019b). For example, the forehead and cheeks are considered prominent regions of interest (ROIs) showing skin color variation. The temporal module allows illustrating the relationships over different frames of a video. EfficientPhys (Liu *et al.*, 2022a) removes preprocessing steps like facial landmark detection, directly processing raw video with 2D convolutional layers. MTTS-CAN (Liu *et al.*, 2020) uses a Temporal Shift Module and averaged frames from adjacent frames to better capture temporal dynamics while reducing noise. Since the rPPG signal represents the variations of blood volume over time, and among the used techniques: Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) are mentioned. After learning relevant spatial features from CNN layers, LSTM and RNN leverage the extracted feature map for modeling temporal dependencies representing the characteristics of the rPPG signal (Yu *et al.*, 2019b; Hu *et al.*, 2019; Niu, Shan, Han & Chen, 2020b). Meta-rPPG (Lee *et al.*, 2020b) employs meta-learning with a 2D CNN and BiLSTM to adapt rapidly to diverse, unlabeled data distributions. However, these approaches have shown less precise results compared to those represented by 3D Convolutional Neural Networks (3D-CNNs), which integrates and encodes spatial and temporal contexts simultaneously (Botina-Monsalve *et al.*, 2022; Perepelkina *et al.*, 2020). 3D-CNNs exhibit the ability to directly extract spatiotemporal features by concentrating solely on the local neighborhood.

While 3D-CNNs are effective at capturing spatiotemporal features within local neighborhoods, they may not sufficiently capture the long-range dependencies essential for accurately estimating vital signs. The kernel-based computations of 3D-CNNs are inherently localized, limiting their ability to grasp global context and extended temporal relationships (Srinivas *et al.*, 2021; Wang, Li, Zhang & Zhang, 2022a). Although stacking multiple convolutional or recurrent layers can, in theory, capture these long-range dependencies, this approach often leads to optimization challenges and increased computational complexity (Wang, Girshick, Gupta & He, 2018). These issues make it difficult to learn long-term correlations among features, which are crucial for understanding physiological characteristics and predicting vital signs accurately.

To address these limitations, transformers have been proposed as a solution for modeling long-range dependencies and global context through self-attention mechanisms. Vision Transformer (ViT) (Dosovitskiy *et al.*, 2020) has demonstrated impressive results in various computer vision tasks by using Multi-Head Self-Attention (MHSA) to capture relationships between distant features, thus improving global context modeling. Transformer-based models for rPPG estimation have shown promise in addressing these challenges. For instance, TransPPG uses a two-stream transformer architecture to separate vital from non-vital features, improving heart rate estimation (Kang *et al.*, 2024). Another approach leverages video transformers to estimate physiological measurements, including heart rate, by modeling spatial and temporal dynamics (Revanur *et al.*, 2022). Similarly, PhysFormer employs a temporal difference transformer to aggregate spatiotemporal features and capture fine-grained temporal skin color differences, enhancing rPPG signal estimation (Yu *et al.*, 2022).

However, ViTs have their own limitations. Unlike CNNs, which possess a strong inductive bias due to their local connectivity and weight sharing, ViTs lack this inherent bias, potentially affecting their generalization, especially with limited data. Moreover, when processing high-resolution images, ViTs tend to demand significantly more computational resources compared to CNNs. Video Vision Transformers (ViViTs) process video data by dividing it into a sequence of patches and applying global self-attention mechanisms sequentially. This approach lacks built-in assumptions, and the absence of a strong inductive bias can lead to weaker generalization performance, particularly on small datasets. Additionally, the computational demands of global self-attention in ViTs escalate when dealing with high-resolution images

To overcome these challenges in rPPG estimation, we propose a novel hybrid approach that synergistically combines 3D-CNNs with MHSA mechanisms in a sequential, feed-forward manner. Our method integrates 3D-CNNs with Video Vision Transformers, leveraging the strengths of both architectures: the strong inductive bias and local feature extraction capabilities of CNNs, and the ability of transformers to model long-range dependencies through global self-attention. In our approach, the 3D-CNNs first process the input data to capture local spatiotemporal variations, generating low-resolution spatiotemporal feature maps. These feature maps are then refined by ViViT's global self-attention mechanisms, which learn long-range dependencies and global contextual relationships. Unlike dual-branch architectures where CNN and transformer branches operate in parallel, our sequential integration ensures progressive refinement, leading to better feature representation, improved accuracy, and computational efficiency. Additionally, this design significantly reduces computational burden because the transformer only processes the refined low-dimensional data.

The blocks within our 3D-CNN architecture are based on Temporal Central Difference Convolution (3DCDC-T) and incorporate spatial-channel attention mechanisms. These blocks are designed to capture temporal gradients and enhance local attention within both spatial and channel dimensions, enriching the spatiotemporal feature representations. The subsequent MHSA layers effectively capture global dependencies, mitigating the limitations of both CNNs and transformers, especially in scenarios with limited data. We have conducted extensive experiments to evaluate the performance of this hybrid model, demonstrating its superiority in estimating rPPG signals and predicting heart rate.

Our main contributions can be summarized as follows:

- Synergistic Feed-Forward Integration of 3D-CNNs with ViViT: We introduce a feed-forward integration where the 3D-CNN captures local spatiotemporal variations, and the MHSA blocks model long-range dependencies. This sequential refinement enables the model to progressively build a more comprehensive representation of the data.
- We evaluated the effectiveness of different facial regions in estimating the rPPG signal and heart rate. Specifically, the model's accuracy was tested using various regions of interest: the entire face, the forehead, the cheeks, and a combination of the forehead and cheeks. Our goal was to identify facial regions with a high density of capillaries while minimizing the interference of noise, thus improving the reliability of rPPG signal extraction.

- We incorporated a Siamese architecture to capture relationships between different facial regions, such as the forehead and cheeks. Additionally, we evaluated a multi-task model capable of estimating both the rPPG signal and heart rate concurrently, optimizing the extraction of both metrics from the input data and ensuring improved performance in vital sign monitoring tasks.

The rest of this paper is as follows. In the second section, state-of-the-art methods used for estimating rPPG signal and heart rate are presented and discussed. In the third section, the overall structure of the hybrid network proposed, in which both models and Siamese network are described in detail. Different experiments are described using various databases, and the results are detailed in Section 4. Lastly, Section 5 provides an overview of the conclusions drawn and outlines future directions.

2.3 Related Works

2.3.1 Traditional Signal Processing Approaches for rPPG

Traditional methods for estimating remote photoplethysmography (rPPG) signals typically follow a two-step procedure. The first step involves face detection, where the task is to detect the face and identify facial regions of interest. Subsequently, in the pixel extraction stage, pixels from the skin areas within the identified facial region are extracted. Following this extraction, a signal processing phase is initiated, wherein data from the red, green, and blue (RGB) light channels is utilized. By analyzing the brightness changes in the RGB channels, we can estimate the rPPG signal, which helps to infer the blood volume pulse (BVP). Numerous studies have refined this method.

For example, (Poh *et al.*, 2010b) employs independent component analysis (ICA) to decompose interrelated RGB signals into three distinct source signals, aiming to isolate the cardiac pulse signal from other incidental signals like noise. The chrominance-based approach in (De Haan & Jeanne, 2013) addresses normalization errors, converting the RGB color space into the chrominance domain to reduce the impact of specular reflections on the skin's surface. Meanwhile, the plane orthogonal to the skin (POS) method in (Wang *et al.*, 2017b) projects the RGB signals onto a plane orthogonal to a normalized skin tone in the RGB space.

While these traditional methods have improved rPPG signal estimation, they remain vulnerable to errors caused by motion artifacts and lighting variations, especially in real-world clinical settings like pediatric intensive care units.

2.3.2 Deep Learning Approaches for rPPG Estimation

To address the limitations of traditional methods, deep learning approaches have been developed to perform end-to-end rPPG signal estimation. These methods leverage convolutional neural networks (CNNs) to extract features directly from video sequences, eliminating the need for separate signal processing stages.

For example, (Chen & McDuff, 2018) uses a motion representation algorithm that relies on a skin reflection model, guided by an attention mechanism, to estimate the heart rate from video frames. Eulerian Video Magnification (EVM) combined with deep learning is explored in (Qiu *et al.*, 2019b), where spatial and temporal features are decomposed and filtered before being fed into a CNN to estimate heart rate. The HR-CNN (Špetlík, Franc & Matas, 2018) includes two parts: a 2D CNN for extracting features and an estimator module for predicting heart rate, though it struggles to learn temporal relationships across frames due to its focus on spatial features.

EfficientPhys (Liu *et al.*, 2022a) simplifies the process by removing complex preprocessing steps like facial landmark detection, using a custom normalization module and 2D convolution layers to efficiently capture spatiotemporal features from raw video. MTTS-CAN (Liu *et al.*, 2020) introduced a Temporal Shift Module (TSM) to capture temporal information efficiently. Instead of raw video frames, the input to their appearance model is an averaged frame from multiple adjacent frames, improving temporal signal capture while reducing noise.

2.3.3 Spatiotemporal Learning with 3D-CNNs

Networks based on 3D-CNNs have shown superior performance over 2D-CNNs for rPPG tasks by learning spatiotemporal features from video sequences. In (Yu *et al.*, 2019d), a 3D spatiotemporal convolutional network, called rPPGNet, was proposed to estimate rPPG signals by improving video input quality and handling compression artifacts. Similarly, PhysNet (Yu *et al.*, 2019b) uses a 3D-CNN architecture to generate spatiotemporal feature representations relevant to rPPG.

In ETA-rPPGNet (Hu *et al.*, 2021a), a Time-Domain Segment Subnet segments the video and uses attention to extract spatial features and aggregate temporal information. The backbone employs 3D convolutions with 1D convolutions to enhance temporal learning and reduce noise. Similarly, SAM-rPPGNet (Hu *et al.*, 2021b) uses a spatial-temporal attention network with strip pooling and a Spatial Attention Module (SAM) to capture facial dynamics and mitigate head motion noise for accurate HR estimation.

2.3.4 Hybrid Models Combining CNNs and RNNs

Several hybrid models have attempted to combine CNNs with Recurrent Neural Networks (RNNs) to improve the capture of temporal dependencies. RhythmNet (Niu *et al.*, 2020b) employs a 2D-CNN to extract spatial features, followed by an RNN (specifically GRU) to model temporal dependencies. While this approach captures temporal information across frames, it introduces significant computational complexity and may not fully capture long-range dependencies.

In (Hu *et al.*, 2019), an end-to-end model using ConvLSTM combined with a 2D-CNN was used to estimate rPPG signals. The CNN extracts spatial features, while ConvLSTM captures temporal dependencies. Meta-rPPG (Lee *et al.*, 2020b) introduces a meta-learner approach to remote heart rate estimation, which rapidly adapts to diverse sample distributions by utilizing unlabeled data. The model employs a 2D CNN combined with a BiLSTM-based spatiotemporal architecture for efficient signal extraction, enhancing its adaptability to new environments. PRNet (Huang *et al.*, 2021a) proposed a one-stage remote heart rate (HR) measurement framework that combines 3D CNN with LSTM, enabling simultaneous spatial and temporal feature extraction for more accurate pulse estimation.

However, PhysNet (Yu *et al.*, 2019b) demonstrated that 3D-CNN architectures generally outperform these hybrid CNN-RNN models by directly integrating spatial and temporal features without the need for complex sequential modeling.

2.3.5 Transformer-Based Models for rPPG Estimation

Transformers have gained significant attention for their ability to model long-range dependencies, initially introduced in (Vaswani *et al.*, 2017) for natural language processing. Vision Transformers (ViT) (Dosovitskiy *et al.*, 2020) extended this concept to image classification, where images are divided into patches, and self-attention is applied to model global dependencies.

TransPPG introduces a two-stream transformer architecture that processes both the foreground (facial regions) and background (environmental noise) in separate streams. By distinguishing between vital and non-vital information, TransPPG improves heart rate estimation in the presence of challenging conditions such as varying lighting and motion artifacts (Kang *et al.*, 2024). In this work (Revanur *et al.*, 2022) they proposed a method that leverages video transformers to estimate instantaneous physiological measurements, including heart rate. The model utilizes transformers to model both spatial and temporal dynamics. PhysFormer (Yu *et al.*, 2022) is an end-to-end video transformer architecture that adaptively aggregates spatiotemporal features for enhanced rPPG signal representation. The key component of PhysFormer is the temporal difference transformer, which enhances quasi-periodic rPPG features with global spatiotemporal attention by focusing on fine-grained temporal skin color differences. TimeSformer (Bertasius *et al.*, 2021) adapted the transformer architecture for video classification by introducing "Divided Space-Time Attention." This method divides attention into temporal and spatial stages, first modeling temporal relationships across frames and then spatial dependencies within each frame. The rPPGTR framework (Zhang, Yang, Yin & Meng, 2023b) applied this attention mechanism to heart rate estimation, fusing local convolutional features with global self-attention to generate a comprehensive rPPG representation. However, hybrid models that fuse local and global features can introduce redundant representations, increasing the computational complexity and overfitting risk.

2.3.6 Hybrid Models Combining 3D-CNNs and Video Vision Transformers

Hybrid models that integrate CNNs with global self-attention mechanisms are particularly effective for rPPG tasks, especially in data-scarce environments. The feature maps produced by 3D-CNNs enable Vision Transformers to learn global relationships more efficiently. This integration is supported by works like rPPGTR (Zhang *et al.*, 2023b), which combines convolution and attention mechanisms for rPPG signal estimation. PhysFormer (Yu *et al.*, 2022) incorporates Temporal Difference Multi-Head Self-Attention (TD-MHSA), where Temporal Difference Convolution (TDC) is applied to query and key projections to capture subtle local temporal variations. This enhances the detection of fine-grained temporal changes in skin color caused by blood flow, allowing the model to capture both local temporal features and global dependencies across frames.

CoAtNet (Dai, Liu, Le & Tan, 2021), vertically stack convolution layers with attention mechanisms, have demonstrated that progressively combining these elements can significantly enhance generalization and computational efficiency. Furthermore, hybrid models like SMAFormer (Zheng *et al.*, 2024), which synergistically fuse multiple attention mechanisms (pixel, channel, spatial), validate the importance of capturing both local and global features to improve segmentation and overall performance. By first generating a low-resolution spatiotemporal map through the 3D-CNN, the Video Vision Transformer processes a more abstract and compressed representation. This reduces computational load while focusing on meaningful global data. Such progressive refinement avoids the redundant feature representations often seen in dual-branch setups, where both branches might struggle to effectively align spatial and temporal information.

However, despite these advancements, existing hybrid models often encounter limitations. Dual-branch architectures, where 3D-CNNs and Vision Transformers process data independently in parallel branches, can lead to redundant computations and increased model complexity. The separation of branches may result in inefficient fusion of spatial and temporal features, making it challenging to align local and global information effectively. Additionally, the redundancy in feature representations can cause overfitting, especially in data-scarce environments, and can hinder real-time application due to increased computational load.

To address these limitations, we propose a synergistic integration of 3D-CNNs followed by a ViViT in a feed-forward manner. This approach offers several advantages over dual-branch architectures. The feed-forward architecture enhances feature consistency, as the global attention mechanisms in the ViViT operate on features already refined by the 3D-CNNs, ensuring coherence across different scales. In contrast, dual-branch architectures often face challenges when fusing features processed separately by CNN and Transformer branches. The hybrid architecture of BoTNet (Srinivas *et al.*, 2021), which replaces convolutional layers in a ResNet bottleneck with multi-head self-attention, demonstrates how this refinement can boost performance by leveraging both local and global dependencies without redundancy. This results in a simpler, more cohesive model design that performs better under limited data conditions and offers improved computational efficiency.

2.4 Proposed Framework

The goal of our proposed framework is to extract physiological information related to blood volume changes, focusing on temporal variations within video frames. By targeting specific facial regions like the forehead and cheek, the model aims to capture subtle skin color variations caused by rhythmic blood flow. These areas, though spatially independent, exhibit correlated temporal signals due to the heartbeat. The proposed framework integrates a 3D-CNNs for capturing local spatiotemporal features and a Vision Transformer for modeling global dependencies, ensuring that local and global patterns are progressively refined in a feed-forward manner.

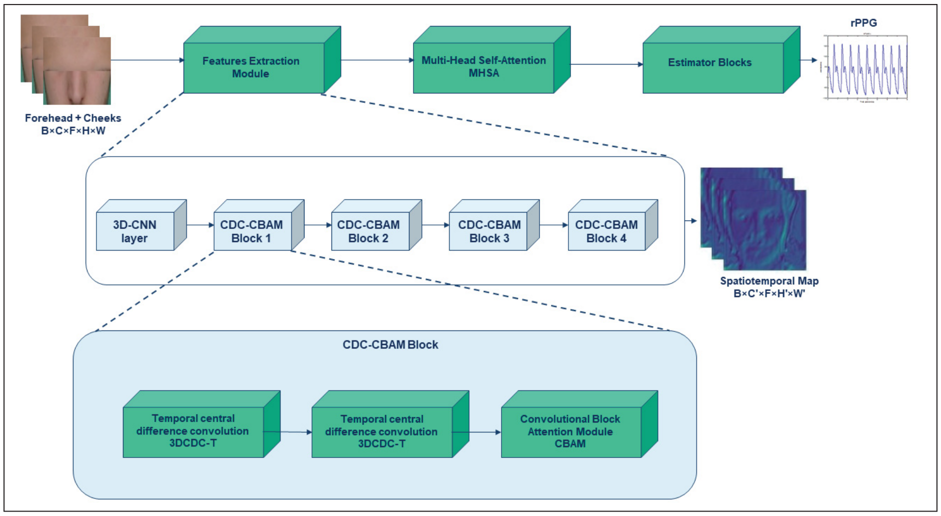

Forehead + Cheeks
B×C×F×H×W
Features Extraction Module
Multi-Head Self-Attention MHSA
Estimator Blocks
rPPG
3D-CNN layer
CDC-CBAM Block 1
CDC-CBAM Block 2
CDC-CBAM Block 3
CDC-CBAM Block 4
Spatiotemporal Map
B×C'×F×H'×W'
CDC-CBAM Block
Temporal central difference convolution 3DCDC-T
Temporal central difference convolution 3DCDC-T
Convolutional Block Attention Module CBAM

Figure 2.1 Overview of the proposed framework: This study investigates different facial regions (full face, forehead, cheek, and concatenated regions). Feature extraction captures local representations, space-time attention models global dependencies, and the estimator predicts the rPPG signal and heart rate

The design of our model, as shown in Fig. 2.1, consists of three key components:

- Feature Extraction Module: Uses 3DCDC-T layers and Convolutional Block Attention Module (CBAM) to capture local spatial and temporal variations while emphasizing informative features that contribute to improved rPPG signal estimation.
- Multi-Head Self-Attention Blocks: Divided space-time attention operates on low-resolution feature maps to model long-range dependencies across frames. This block facilitates learning global contextual relationships by analyzing spatial and temporal dimensions.
- Estimator Blocks: Acts as the decoder module within the CNN architecture. Its primary role is to predict the rPPG signal and/or heart rate (HR) by using the refined features from the self-attention blocks to make accurate predictions.

By integrating these components in a feed-forward manner, the model first captures local spatial and temporal features with the 3D-CNNs, then refines them through global attention. The following subsections describe each phase in detail.

2.4.1 Feature Extraction Module

The feature extraction module is composed of a designed structure that comprises three cascaded blocks based on 3D convolutional neural network (3D-CNN) layers, inspired by this work (Liu, Wei, Kuang & Ma, 2022b). This architecture is depicted in Fig. 2.1. Each of these blocks consists of two consecutive layers of temporal central difference convolution (3DCDC-T), seamlessly followed by a fusion of spatial and channel attention mechanisms. A downsampling operation within each block reduces the spatial dimensions while increasing the channel dimension and keeping the depth the same. The 3DCDC-T blocks are designed to capture temporal correlations of gradients, allowing the model to detect subtle skin color variations caused by blood flow, while the spatial-channel attention mechanisms enhance the model's ability to focus on relevant regions and channels. This approach enables the construction of enriched spatiotemporal feature maps, crucial for accurate rPPG signal estimation. With the aim to capture representations pertinent to the characteristics of the rPPG signal, the initial layer is a standard 3D convolutional neural network (3D-CNN). This initial layer aids in extracting more suitable representations that are closely related to physiological parameters.

2.4.1.1 3D Central Difference Convolutional Network

The rPPG signal estimation relies on detecting subtle changes in skin color caused by blood flow. Therefore, the model should focus on the time derivative to extract relevant temporal features from consecutive frames. 3D-Central difference convolutional (3D CDC) network is used to capture changes in intensity across neighboring pixels that happened across spatiotemporal features. Hence, temporal central difference convolution (3DCDC-T) is specifically designed to detect local temporal variations between successive frames by calculating the differences in pixel intensities over time. 3DCDC-T extends standard convolutional neural networks from the spatial domain to temporal dimension by convolving the input videos with another filter based on central difference (Yu *et al.*, 2021; Zhao *et al.*, 2021; Liu *et al.*, 2022b). Temporal Central Difference Convolution (3DCDC-T) based on two separated local receptive regions R'

representing the current time step and R'' gives information in the adjacent time steps :

$$\begin{aligned} 3DCDC_T(r_0) &= \sum_{r_n \in C} w(r_n).x(r_0 + r_n) \\ &+\theta(-x(r_0).\sum_{r_n \in R''} w(r_n)) \end{aligned} \tag{2.1}$$

where r_0 denotes current position, local receptive regions represented by C and r_n enumerates locations in these regions. Integrating 3DCDC-T with a standard convolution neural network enables the incorporation of temporal gradient correlations into the feature map. This allows the rPPG model to leverage these supplementary features, enabling a sharper emphasis on local temporal variations over time. Consequently, unwanted noise interference becomes more effectively controlled and regulated. Implementing 3DCDC-T directs the network's attention to temporal shifts caused by subtle variations in intensity, which are associated with changes in blood volume.

2.4.1.2 Spatial and Channel Attention Mechanism

Blood flow is more detectable in certain facial regions, notably the forehead and cheeks, compared to other areas. Consequently, the network has to emphasize the importance of features within those regions while filtering out pixels with constant values, which are regarded as noise in the context of rPPG signal recovery. The attention mechanism guides models on which spatial and channel features should be emphasized and which can be disregarded for the accurate estimation of the rPPG signal. The Convolutional Block Attention Module (CBAM) is employed as this attention mechanism. It operates on the feature map generated by the convolutional network. CBAM is divided into two blocks: spatial and channel-wise attention.

The spatial attention module aims to illustrate spatial correlation within channels in the feature map. It prioritizes the most consistent information and ignores features considered as noise to the network. The spatial attention module comprises two steps: (1) applying 3D average pooling; and (2) performing 3D max pooling operations over the channel and temporal dimensions. The obtained attention weights are multiplied element-wise with the original feature map to adjust the spatial map (Woo, Park, Lee & Kweon, 2018; Liu *et al.*, 2022b).

The channel attention module models the relation between channels by capturing the necessary channels needed for rPPG signals. Similar to the spatial attention module, it performs 3D average pooling and 3D max pooling operations but over the spatial and temporal dimensions. By computing both channel-wise and spatial-wise attention, the spatial and channel attention mechanism aids the network in prioritizing pertinent spatial regions and channels, ensuring that the most informative features are captured for accurate rPPG estimation.

2.4.2 Multi-Head Self-Attention Blocks

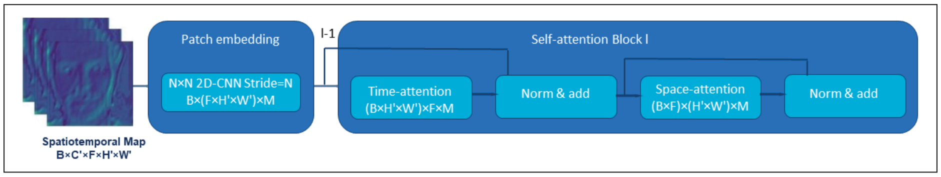


Figure 2.2 Video Vision Transformer module takes as input feature map obtained with the feature extraction network then transformed into a patch embedding. The patch embeddings are then fed into the divided time-space attention mechanism

The MHSA blocks is used to aggregate the encoded features within the spatial and temporal feature map acquired through the 3D-CNN blocks. TimeSformer, which encompasses self-attention across both spatial and temporal dimensions, is employed to characterize temporal and spatial features derived from a sequence of frames. This approach effectively illustrates long-range dependencies through the utilization of divided space-time attention (Bertasius *et al.*, 2021). These divided space-time attention blocks serve as a means to capture and portray global temporal and spatial relationships embedded within the rich semantic information extracted by the 3D-CNN blocks. Within each dividesd space-time attention block separate computations for temporal and spatial attention are performed independently.

The feature map obtained from the first module will be fed into the divided space-time attention. To prepare the feature map for this module, each frame within the map is transformed into a patch embedding of size $N \times N$ using non-overlapping 2D convolutional neural network (2D-CNN), as shown in Fig. 2.2. The kernel size and stride of the 2D-CNN are set to match the patch size. The resulting patches are flattened into a vector of size m, representing the embedding dimension. We apply time attention then space attention. Hence for temporal attention we rearrange the features from $R^{b\times(h\times w\times f)\times m}$ to $R^{(b\times h\times w)\times f\times m}$, and for spatial attention the output features are rearranged to $R^{(b\times f)\times(h\times w)\times m}$. This module contains 6 encoding blocks. For each block i and over multiple attention heads, we compute the query, key, and value using the encoded representation $x_{(n,f)}$ obtained from the previous block :

$$query_{(n,f)} = W_{query} LayerNorm(x^{(i-1)}_{(n,f)}) \tag{2.2}$$

$$key_{(n,f)} = W_{key} LayerNorm(x^{(i-1)}_{(n,f)}) \tag{2.3}$$

$$value_{(n,f)} = W_{value} LayerNorm(x^{(i-1)}_{(n,f)}) \tag{2.4}$$

where n represents the patch index, f is an index for frames, i the block index and $W_{query}, W_{key}, W_{value}$ are learnable matrix for query, key, and value.

2.4.3 Estimator Block

The output of the MHSA module is passed through an estimator block. It is specifically designed to learn the intricate mapping process between the extracted features and the rPPG signal. This block consists of two layers of 3D convolutional neural networks (CNNs), each followed by an ELU activation function, to interpret the encoded features effectively. Afterward, 3D adaptive average pooling captures the relevant spatial-temporal patterns. A final 3D-CNN layer is used to predict the rPPG signal based on the encoded features. Given the correlation between rPPG signals and heart rate over time, we explored a multi-tasking network with two branches—one for rPPG estimation and the other for heart rate prediction, eliminating the need for pre-processing to predict heart rate. The heart rate estimator includes a 3D-CNN followed by a fully connected layer to predict the average heart rate. By incorporating multitasking, the network can leverage shared features for both tasks.

2.4.4 Siamese Network

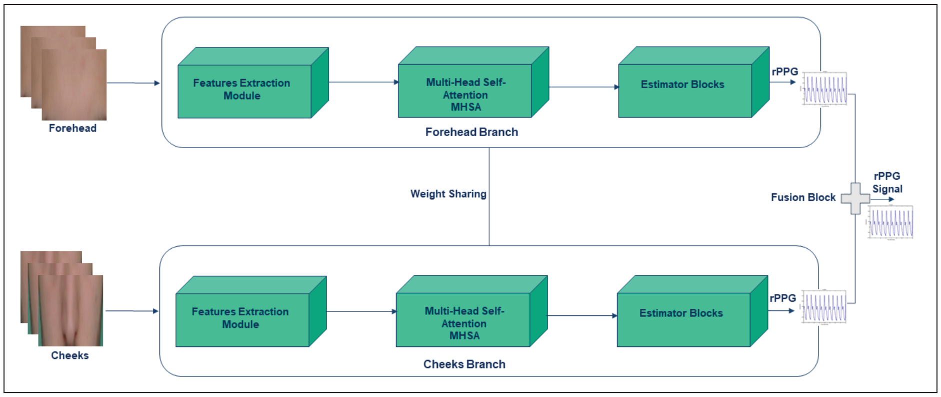


Figure 2.3 Siamese network approach

We propose to integrate the Siamese network into our approach to evaluate the impact of learning correlated features from various regions of interest. A Siamese network is a deep learning architecture that has two identical networks with shared weights. This design allows the model to effectively learn common features from different perspectives. In our case, we deployed a Siamese network where the first branch takes input from the forehead, while the second branch focuses on the cheek region (Tsou, Lee, Hsu & Chang, 2020). Each branch adopts the architecture of the network described earlier. This enables the model to capture interconnected physiological characteristics from different regions, playing a crucial role in accurately estimating remote photoplethysmography (rPPG) signals simultaneously. After estimating the rPPG signals independently in each branch, we incorporate a fusion step by adding together the outputs from both branches, as shown in Fig. 2.3. This fusion process serves the purpose of arriving at a conclusive decision by taking into account not only the individual outputs of each branch but also the relationship or correlation between these outputs.

2.5 Experimental Protocol

2.5.1 Datasets

1. Dataset UBFC-rPPG: It was published in 2017 (Bobbia, Macwan, Benezeth, Mansouri & Dubois, 2019); it contains 48 videos of 48 subjects recorded with webcam. Each video is almost 2 minutes long, with 30 frames per second and a spatial resolution of 640 × 480 pixels. The videos are synchronized with a finger pulse oximeter CMS50E which is used to get the ground truth (PPG values and heart rate) of each recorded subject. In all videos, subjects were sitting one-meter away from the camera; in order to change heart rate they were required to play a time sensitive mathematical game during the recording.
2. Dataset VicarPPG-2: It was released in 2020 (Gudi, Bittner & van Gemert, 2020) in order to evaluate algorithms to estimate heart rate and heart rate variability from 40 different videos. It contains videos of 10 subjects, where they were asked to sit one meter in front of webcam with frame rate of 60 fps. The ground truth of each subject is extracted using ECG board to get ECG signals also with pulse oximeter device to get PPG signals. In each video, the participant is recorded under 4 different conditions: baseline conditions, playing a difficult Stroop Task related game, performing multiple different head movements and cooling down after exercise. In the movement condition, subjects performed the eight different movements at different times during the recording.
3. Dataset ECG-Fitness: It presents a challenge in remote heart rate (HR) measurement tasks (Špetlík *et al.*, 2018), focusing on individuals engaged in physical activities on fitness machines across varying lighting conditions. This dataset encompasses 17 subjects (14 male, 3 female) aged between 20 and 53 years, participating in activities like speaking, rowing, and using stationary bikes and elliptical trainers. Video recordings were captured using web cameras, with one camera on a stable tripod, resulting in 204 videos depicting fitness-related actions. The recordings were conducted under three distinct lighting setups: natural light, 400W halogen light, and 30W LED light. Concurrently, the subjects' heart rates were monitored via electrocardiogram (ECG) recordings, enabling precise HR computation through Python algorithms.

2.5.2 Loss function

In our proposed network architecture, we applied two different approaches: first the model is trained by only estimating rPPG signal and using as ground truth PPG signal. In this case the model try to learn a signal similar to the reference one which is measured using the pulse oximeter. The model should focus more on the trend of both signals and should ignore the amplitude of waveform. This is achieved by measuring the relationship between the predicted signal and the ground truth with Negative Pearson Correlation loss function :

$$Loss = 1 - \frac{\sum_{i=1}^{T}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{T}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}. \tag{2.5}$$

where T is number of frames per video clip, y_i is the predicted rPPG signals, $\hat{y}_i$ represents the ground truth PPG signal, $\bar{y}$ and $\bar{\hat{y}}$ are the mean of the predicted and ground truth values.

In the second approach, as suggested in (Zhang *et al.*, 2023b), the model predicts simultaneously the rPPG signal and the average heart rate by combining both losses: the rPPG signal loss and the heart rate loss; this improves the performance by guiding the model to learn the relationship between the rPPG waveform and the average heart rate. The mean Absolute Error (MAE) was applied as the heart rate loss as following:

$$Loss = \frac{1}{n}\sum_{i=1}^{n}|HR(i) - \hat{H}R(i)|. \tag{2.6}$$

where HR represents the heart rate value and $\hat{H}R$ is the ground truth of the heart rate.

2.5.3 Training details

Prior to training, a Region of Interest (ROI) was defined using the Mediapipe tool (Google Developers, 2024), which applies facial landmark detection to isolate key areas of the face. For each video clip, facial landmarks were detected across 128 frames, and four distinct facial regions were extracted: the entire face, forehead, cheeks, and a combination of the forehead and cheeks. To standardize the input size, all extracted regions were resized to 96×96 pixels.

Each video segment was synchronized with the corresponding Photoplethysmography (PPG) signal, where the heart rate label represented the average heart rate over that period. In the feature extraction module, the channel dimensions were progressively expanded to 6, 16, 24, 36, while the temporal depth remained fixed at 128, reflecting the number of frames in the sequence. Multi-Head Self-Attention (MHSA) blocks were used to capture long-range dependencies, operating on patches of size 16×16. The attention mechanism employed 12 attention heads with a depth of 12 and an embedding dimension of 768.

The model was first pre-trained on the VicarPPG-2 dataset. A sliding window of 128 non-overlapping frames was used, resulting in a total of 4480 samples for pre-training. For fine-tuning, we utilized the UBFC-RPPG database. In alignment with the methodology described in (Song *et al.*, 2021), we generated 1176 samples, with the first 36 subjects used for training and the remaining 12 subjects for testing.

The training was carried out using the Rectified Adam (RAdam) optimizer (Liu *et al.*, 2019), with a learning rate of 0.0005. Training was conducted for 150 epochs during pre-training and 50 epochs for fine-tuning, with a batch size of 4. After each convolutional layer, an activation function was applied to introduce non-linearity and enhance model performance. All experiments were implemented in PyTorch (Paszke *et al.*, 2019), and the network was trained on a single NVIDIA A100SXM4 GPU.

2.5.4 Cross-database Testing

To evaluate the generalization capability of our proposed framework, we performed cross-database testing by training the model on the UBFC-rPPG and VicarPPG-2 datasets and then testing it on the ECG-Fitness dataset. This approach allows us to assess how well the model can generalize to unseen data across different conditions and environments.

Since the ECG-Fitness dataset provides only heart rate labels derived from ECG signals, the testing focuses solely on heart rate (HR)-related metrics. We did not assess the Pearson Correlation Coefficient of the signal because the dataset lacks ground truth data for rPPG signals.The purpose of cross-database testing is to verify that the model, trained on datasets with diverse conditions, such as lighting, skin tone, and motion, can effectively predict heart rate across datasets with different characteristics.

2.5.5 Evaluation metrics

To evaluate the model, four performance metrics were applied: mean absolute error (MAE), root mean squared error (RMSE), mean percentage error (MPE) and Pearson's correlation coefficient (R):

1. Mean absolute error MAE: It is calculated by taking the average absolute difference between the predicted values $HR_{predict}$ and the ground truth values $HR_{reference}$:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|HR^{i}_{predict} - HR^{i}_{reference}|. \tag{2.7}$$

2. Root mean squared error RMSE: It measures the average magnitude difference between the predicted value of heart rate and the actual one

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(HR^{i}_{predict} - HR^{i}_{reference})^{2}}. \tag{2.8}$$

3. Mean percentage error MPE: MPE represents the average percentage errors how much the predicted value differ from the actual value of heart rate:

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{HR^{i}_{predict} - HR^{i}_{reference}}{HR^{i}_{predict}}\right) \times 100. \tag{2.9}$$

4. The Pearson correlation coefficient R : It measures the linear correlation between the two signals $rPPG_{predict}$ and $PPG_{reference}$:

$$R = \frac{\mathrm{cov}(rPPG_{predict}, PPG_{reference})}{\mathrm{std}(rPPG_{predict}) \cdot \mathrm{std}(PPG_{reference})}. \tag{2.10}$$

2.5.6 Ablation Study

Many experiments were conducted to study the impact of the different approaches on the accuracy of rPPG estimation model. First, we evaluated the impact of integrating MHSA within a 3D-CNNs architecture. This experiment aimed to examine how the combination of 3D-CNNs and ViViT can enhance model generalization and increase its capacity. Specifically, we focused on three key aspects to assess how this hybrid approach refines feature learning by capturing both local and global spatiotemporal dependencies, leading to improved performance in tasks such as rPPG signal estimation.

- Training a model using only 3D-CNN blocks while excluding MHSA blocks.
- Evaluating an approach constructed only with ViViT (Divided Space-Time Attention).
- Combining both the 3D-CNNs model and MHSA in a hybrid approach.

To investigate the effect of Temporal Central Difference Convolution (3DCDC-T) and the Convolutional Block Attention Module (CBAM) on constructing an abstract spatiotemporal map that captures local variations in skin color, we conducted a series of comparative experiments. In the first configuration, we used standard 3D convolution without CBAM as a baseline. We then explored the impact of adding CBAM to standard 3D convolutions to assess the role of spatial and channel attention. Finally, we replaced the standard 3D convolution with 3DCDC-T, without CBAM, to understand the contribution of temporal feature extraction. In each case, the spatiotemporal map was passed to the MHSA module to evaluate how these modifications affected feature extraction and overall model performance.

Furthermore, the model is trained using two different datasets: UBFC and Vicar-PPG. The aim was to investigate the impact of various regions of interest (RoIs) on the model's ability to accurately estimate the PPG signal within each dataset. In each experimental setup, different RoIs are used as input to the model, encompassing the entire face region, forehead area, cheeks region, and a combination of the forehead and cheeks regions concatenated together. The network systematically employed these diverse RoIs as input data to identify the most relevant regions that could enhance the model's prediction.

Moreover, the Siamese network is implemented to explore the architecture capability in enhancing the accuracy of rPPG signal estimation. In this approach, we utilized the forehead and cheeks regions as inputs for each network branch. This architecture enabled the model to simultaneously learn correlated features from both regions and leverage their complementary information.

Additionally, a multitasking model is introduced, where the model predicts both the rPPG signal and the average heart rate along that period. This methodology enables the model to capture the intricate relationship between these two crucial vital signs. This improves the model's performance by helping it better understand how the rPPG waveform and the average heart rate are connected.

2.6 Results and Discussion

2.6.1 Comparison with State-of-the-Art Methods on the UBFC-rPPG Dataset

The comparative results of our proposed method against several state-of-the-art approaches on the UBFC-rPPG dataset are presented in Table 2.1. As shown, deep learning-based techniques substantially outperform traditional methods such as ICA, CHROM, and POS, all of which exhibit significantly higher Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values. For instance, the POS method reports an MAE of 6.52, while our model achieves a much lower MAE of 0.79, indicating a marked improvement in accuracy.

Among the deep learning models, our approach delivers the best performance, with an MAE of 0.79 and an RMSE of 0.80. This surpasses well-established models such as PhysFormer, EfficientPhys, and TransPPG, which report MAE values of 1.02, 1.14, and 1.09, respectively.

The improved results can be attributed to the hybrid architecture of 3D-CNNs and MHSA, which effectively captures both local spatiotemporal features and global dependencies across frames. Additionally, the integration of 3DCDC-T and CBAM further refines the representation of subtle skin color variations, improving the overall rPPG signal estimation.

Furthermore, in terms of correlation (R), our model achieves an impressive value of 0.99, indicating a strong alignment between the predicted and ground-truth heart rates. This high correlation emphasizes the robustness and reliability of our model in estimating heart rates under controlled conditions, particularly on the UBFC dataset.

Table 2.1 Comparison among different methods on the UBFC-rPPG dataset

Method	**MAE**	**RMSE**	**R**
POS(Wang *et al.*, 2017b)	6.52	10.52	0.86
CHROM(De Haan & Jeanne, 2013)	6.37	9.10	0.91
ICA(Poh *et al.*, 2010b)	5.17	11.76	0.65
META-rPPG(Lee *et al.*, 2020b)	5.97	7.42	0.53
EVM-CNN(Qiu *et al.*, 2019b)	4.90	5.89	0.64
DeepPhys(Chen & McDuff, 2018)	2.90	3.63	0.98
HeartTrack(Perepelkina *et al.*, 2020)	2.41	3.37	0.98
ETA-rPPGNet(Hu *et al.*, 2021c)	1.46	3.97	0.95
EfficientPhys(Liu *et al.*, 2022a)	1.14	1.81	0.99
TransPPG(Kang *et al.*, 2024)	1.09	3.05	0.93
Physformer(Yu *et al.*, 2022)	1.02	2.81	0.96
Our	**0.79**	**0.80**	**0.99**

2.6.2 Cross-Dataset Evaluation on ECG-Fitness

To further evaluate the generalization capability of our proposed model, we conducted cross-dataset testing on the ECG-Fitness dataset. The results, as shown in Table 2.2, highlight the effectiveness of our approach in a more challenging scenario where only heart rate labels derived from ECG signals are available, without direct ground truth rPPG signals.

Our model achieves a significantly lower MAE of 6.21, RMSE of 13.71, and MPE of 6.01, outperforming both traditional methods like ICA and POS, which report much higher MAE values of 25.17 and 23.46, respectively. This demonstrates the limitations of traditional approaches in handling more complex and varied datasets, such as ECG-Fitness, which introduces greater variability in conditions like lighting and motion. Among the deep learning methods, our approach also surpasses models like RhythmNet, HR-CNN, and rPPGTR, further proving its robustness and ability to generalize well to different datasets.

Table 2.2 Comparison among different methods on the ECG-Fitness dataset

Method	**MAE**	**RMSE**	**MPE**
ICA(Poh *et al.*, 2010b)	25.17	29.64	32.62
CHROM(De Haan & Jeanne, 2013)	21.37	33.47	-
POS(Wang *et al.*, 2017b)	23.46	30.60	31.51
RhythmNet(Niu *et al.*, 2020b)	16.82	20.47	-
HR-CNN(Špetlík *et al.*, 2018)	14.48	19.15	20.83
DeepPhys(Chen & McDuff, 2018)	11.62	19.67	16.33
rPPGTR(Yu *et al.*, 2019d)	10.31	15.21	13.98
Our	**6.21**	**13.71**	**6.01**

2.6.3 RoI Performance Comparison

To determine the most effective region of interest (RoI) for accurate blood volume pulse estimation, we tested several regions using both the UBFC and VicarPPG-2 datasets. Initially, the model was trained using the entire face as the input, followed by experiments where the forehead and cheeks were tested independently. Lastly, we evaluated the combined forehead and cheek regions as a single input.

The results, presented in Tables 2.3 and 2.4, show the performance metrics for each RoI across both datasets. Across both datasets, using the combined forehead and cheeks region consistently led to improved performance metrics compared to using the entire face, forehead, or cheeks alone. This is likely due to the model's ability to eliminate shiny spots and focus more accurately on areas where skin color changes caused by blood flow are most detectable. Figure 2.4 illustrates these findings, with the Bland–Altman analysis showing a more stable distribution of heart rate estimates when using the combined RoI.

Table 2.3 The performance of different regions on the UBFC-rPPG dataset

Region	**MAE**	**RMSE**	**MPE**	**R**
Face	1.78	4.90	3.01	0.91
Forehead	0.91	1.63	0.81	0.99
Cheeks	0.98	1.90	1.10	0.96
Forehead + Cheeks	**0.79**	**0.80**	**0.51**	**0.99**

Table 2.4 The performance of different regions on the VicarPPG-2 dataset

Region	**MAE**	**RMSE**	**MPE**	**R**
Face	2.53	2.88	3.51	0.85
Forehead	3.96	4.58	4.70	0.91
Cheeks	1.73	1.96	2.03	0.95
Forehead + Cheeks	**0.77**	**1.31**	**1.00**	**0.98**

2.6.4 Siamese network

We implemented a Siamese network where the forehead region served as the input for the first branch and the cheek region for the second. This architecture aims to capture correlated features from both regions to improve the rPPG estimation process. After estimating the rPPG signals from each branch, the outputs were fused by performing an addition operation between them. The experimental results for the Siamese network are presented in Table 2.5. Compared to the individual and combined RoIs tested in Tables 2.3 and 2.4, the Siamese network shows less favorable performance. For the UBFC dataset, it achieved an MAE of 5.2 and RMSE of 7.82, while on VicarPPG-2, the MAE was 3.7, and RMSE was 5.98. This drop in performance suggests that the Siamese network struggles to learn a unified representation from both regions, possibly due to difficulties in capturing the temporal correlations between the spatial variables.

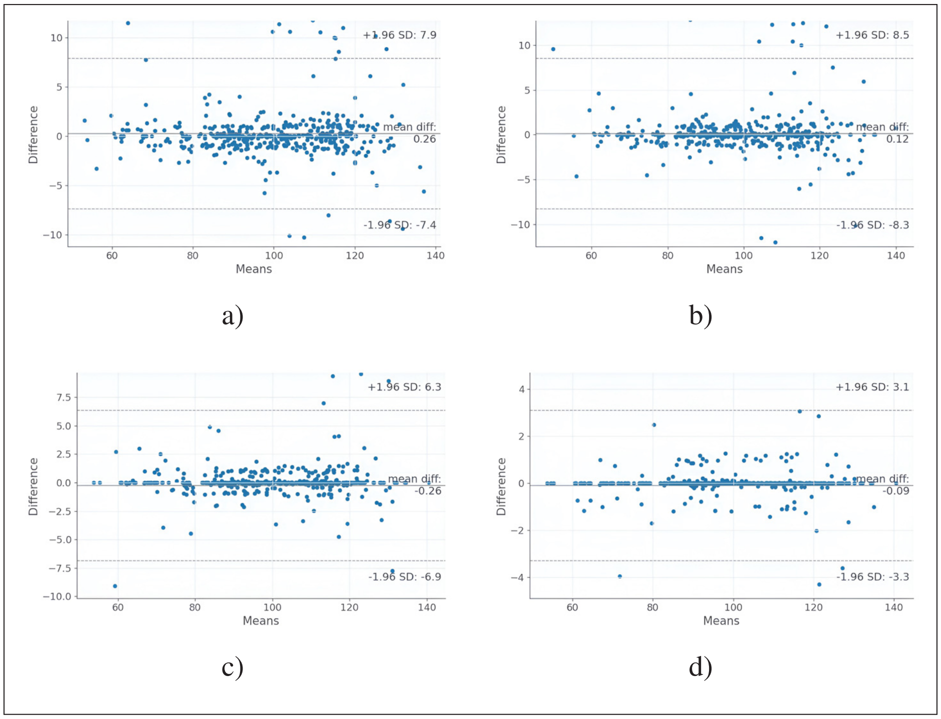


Figure 2.4 Bland–Altman plots of different regions of interest RoIs. (a) Face. (b) Cheeks. (c) Forehead. (d) Forehead and cheeks

Table 2.5 Experimental results of the Siamese network

Dataset	**MAE**	**RMSE**	**MPE**	**R**
UBFC	5.20	7.82	2.00	0.91
VicarPPG-2	3.70	5.98	10.00	0.93

2.6.5 Multitasking vs Single Task

Furthermore, We evaluated the performance of a multitasking architecture, comparing it to single-task approaches for rPPG and heart rate estimation. In the multitasking setup, negative Pearson correlation was used for rPPG and mean absolute error (MAE) for heart rate, with optimized weights for both. As shown in Table 2.6, rPPG estimation alone performed better, achieving an MAE of 0.93 and RMSE of 1.89, while heart rate prediction alone resulted in an

MAE of 4.37 and RMSE of 9.77. These results highlight the multitasking model's ability to capture relationships more effectively, particularly for rPPG.

Table 2.6 Impact of multitasking and single-task learning

Model	**MAE**	**RMSE**	**MPE**	**R**
HR only	4.37	9.77	5.24	–
rPPG only	0.93	1.89	0.80	0.94

2.6.6 Impact of Integrating MHSA into 3D-CNNs

We examined the impact of incorporating Multi-Head Self-Attention (MHSA) into the 3D-CNN architecture. This hybrid model capitalizes on the efficient spatiotemporal feature extraction of 3D-CNNs, while MHSA captures long-range dependencies across frames. The combination of these two mechanisms improved both generalization and model capacity. As shown in Table 2.7, the hybrid model significantly outperforms the 3D-CNN-only and ViViT-only models, achieving superior accuracy. We also assessed the performance of models trained with only 3D-CNN or MHSA. The standalone MHSA model performed poorly, indicating that without the inductive bias provided by CNNs, MHSA struggles to capture relevant features effectively, particularly with smaller datasets. Integrating 3D-CNNs into the vision transformer framework expedited optimization and increased the model's ability to generalize across various datasets. This demonstrates the complementary nature of 3D-CNNs and MHSA in boosting the overall performance of the model.

Table 2.7 Impact of integrating MHSA into the 3D-CNNs architecture

Architecture	**MAE**	**RMSE**	**MPE**	**R**
3D-CNNs only	2.65	2.90	3.50	0.74
ViViT only	5.76	3.39	7.60	0.44
Hybrid Model	**0.79**	**0.80**	**0.51**	**0.99**

2.6.7 Impact of 3DCDC-T and CBAM on rPPG Estimation

The results in Table 2.8 highlight the significant improvements gained by integrating 3DCDC-T and CBAM into the model. The inclusion of CBAM in a standard 3D convolution architecture reduces both MAE and RMSE, indicating enhanced attention to the most relevant spatial features. Furthermore, replacing the standard 3D convolution with 3DCDC-T further decreases error rates, demonstrating the effectiveness of 3DCDC-T in capturing important temporal variations that are critical for rPPG signal estimation. These results confirm that the combination of CBAM and 3DCDC-T substantially improves feature extraction, leading to more accurate and robust rPPG signal predictions.

Table 2.8 Impact of 3DCDC-T and CBAM on rPPG signal estimation

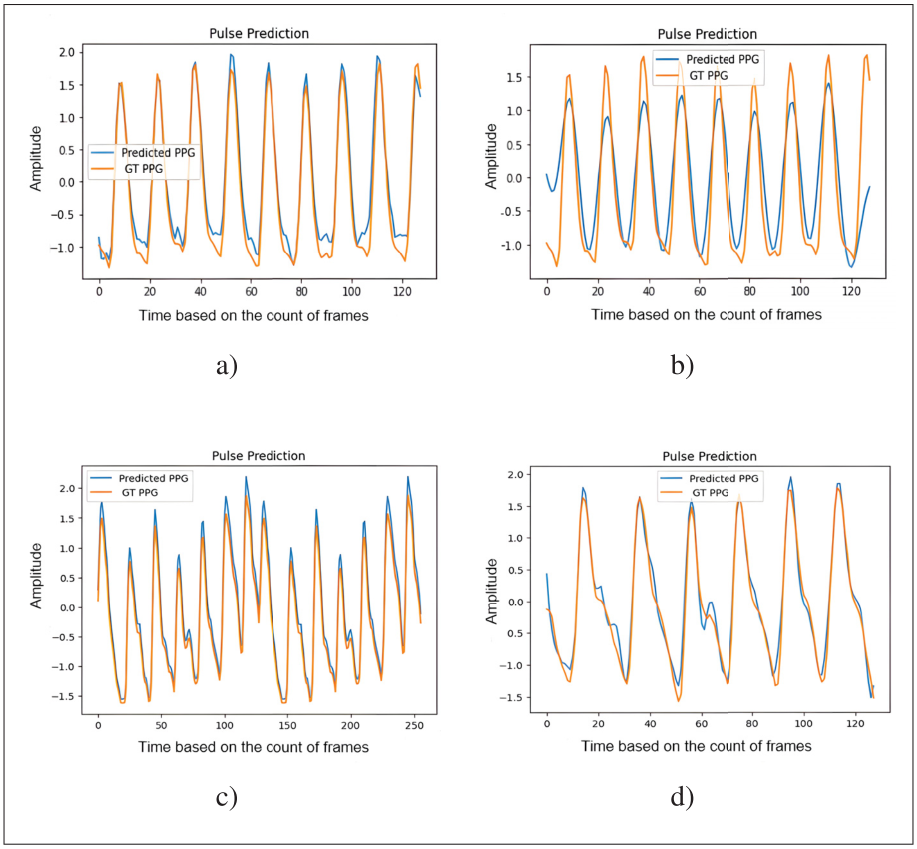
Approach	**MAE**	**RMSE**
3D Convolution (without CBAM)	1.37	2.17
3D Convolution (with CBAM)	1.24	1.79
3DCDC-T (without CBAM)	0.92	1.35

2.6.8 Discussion

Figure 2.5 visually illustrates the alignment between the estimated rPPG signals and the actual ground truth. The waveform patterns of the estimated signals closely resemble the real signals, demonstrating significant correlation, especially in the shape of the waveform. This is key to calculating heart rate using the power spectrum of the estimated rPPG signal.

2.6.8.1 Hybrid Model with 3D-CNNs and Global Self-Attention for rPPG Estimation

In our hybrid model, which integrates 3D Convolutional Neural Networks (3D-CNNs) with global self-attention, we effectively combine these mechanisms to enhance rPPG signal estimation. The 3D-CNNs capture local spatiotemporal features, such as subtle skin color variations due to blood flow, which are critical for accurate rPPG estimation. However, traditional CNNs



a)

b)

c)

d)

Figure 2.5 Illustration of estimated rPPG signals and ground truth within: (a), (b) VicarPPG-2 and (c), (d) UBFC dataset

struggle to capture long-range dependencies across frames. To address this, global self-attention models global contextual relationships through self-attention, enabling the capture of long-range dependencies between frames and spatial regions. This combination is particularly valuable for rPPG tasks, where physiological changes occur over time and across spatial regions. As a result, the hybrid model demonstrates improvements in robustness to noise, lighting variations, and motion artifacts, outperforming traditional CNN-based approaches, particularly in cross-dataset validation on the ECG-Fitness dataset.

2.6.8.2 Impact of Incorporating MHSA into 3D-CNN Architecture

The integration of Multi-Head Self-Attention (MHSA) into the 3D-CNN architecture significantly enhances the model's ability to capture complex temporal and spatial dependencies in video data. While 3D-CNNs process spatiotemporal information effectively, they are limited by their receptive fields. MHSA overcomes this by allowing the network to focus on multiple input regions simultaneously, leading to more robust feature representations. This results in improved accuracy for dynamic tasks like heart rate estimation from videos. The MHSA mechanism, coupled with 3D-CNNs, strengthens the model's generalization ability, as shown in cross-dataset testing.

2.6.8.3 Efficiency and Real-Time Application

In addition to improved accuracy, the feed-forward integration of 3D-CNNs and MHSA ensures computational efficiency. By reducing spatial resolution through 3D-CNN layers, the MHSA operates on lower-dimensional feature maps, minimizing redundancy. This design reduces the computational load without sacrificing accuracy, enabling the model to perform real-time rPPG estimation with an inference time of 0.22 seconds per video. The model's adaptability to small datasets and its streamlined architecture make it suitable for real-time applications like remote health monitoring.

2.6.8.4 Robustness with Small Datasets

The results from both the UBFC-rPPG and ECG-Fitness datasets demonstrate the superior performance of our approach. Our hybrid model leverages the strong inductive bias of 3D-CNNs for local feature extraction and the Video Vision Transformer's (ViViT) ability to capture global dependencies, addressing the challenges posed by small datasets. Additionally, the sliding window technique, using 128-frame segments, effectively increases training data and helps the model learn from various parts of the video, enhancing robustness. This makes our approach well-suited for real-time remote heart rate estimation, where data is often limited.

2.7 Conclusion

In the context of the COVID-19 pandemic, the importance of Remote Photoplethysmography (rPPG) has increased significantly as a non-invasive solution for monitoring vital signs, particularly heart rate, by analyzing subtle changes in skin color. This technology enables continuous, contact-free patient monitoring, which is particularly beneficial for healthcare providers, especially when dealing with vulnerable populations such as pediatric patients. By eliminating the need for physical contact, rPPG facilitates better decision-making and uninterrupted observation of patient health status.

In this study, we developed a hybrid model that combines the strong inductive bias of 3D Convolutional Neural Networks (3D-CNN) with the model capacity of Multi-Head Self-Attention (MHSA), enabling the capture of both local and global spatiotemporal features. This integration has shown to enhance generalization and model capacity, leading to improved accuracy in rPPG signal and heart rate estimation from RGB videos. The model was trained using various regions of interest, including the face, forehead, cheeks, and a combination of forehead and cheeks. A Siamese network architecture was employed to investigate the effect of using multiple facial regions for more precise estimation. Additionally, we explored multitasking scenarios for jointly predicting rPPG signals and heart rates, alongside independent estimations for each metric. These analyses demonstrate the efficacy of our approach in advancing remote heart rate estimation techniques, providing valuable insights for real-world healthcare applications.

Future work will focus on applying our rPPG estimation method to a video dataset of pediatric patients in intensive care units, in order to validate the algorithm in clinical conditions and further assess its practical utility in healthcare settings (Boivin *et al.*, 2023).

CHAPTER 3

PICU FACE AND THORACOABDOMINAL DETECTION USING SELF-SUPERVISED DIVIDED SPACE–TIME MAMBA

Mohamed Khalil Ben Salah[1] , Philippe Jouvet[2] , Rita Noumeir[1]

[1] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[2] Research Center at CHU Sainte-Justine Hospital, University of Montreal,
3175 Chem. de la Côte-Sainte-Catherine, Montréal, Québec, Canada H3T 1C5

Article published in « Life Special Issue: Current and Future Perspectives of Artificial Intelligence in Medicine » in October 2025.

3.1 Abstract

Non-contact vital sign monitoring in Pediatric Intensive Care Units is challenged by frequent occlusions, data scarcity, and the need for temporally stable anatomical tracking to extract reliable physiological signals. Traditional detectors produce unstable tracking, while video transformers are too computationally intensive for deployment on resource-limited clinical hardware. We introduce Divided Space-Time Mamba, an architecture that decouples spatial and temporal feature learning using State Space Models to achieve linear-time complexity, over 92% lower than standard transformers. To handle data scarcity, we employ self-supervised pre-training with masked autoencoders on over 50k domain-specific video clips and further enhance robustness with multimodal RGB-D input. Our model demonstrates superior performance, achieving 0.96 mAP@0.5, 0.62 mAP50-95, and 0.95 rotated IoU. Operating at 23 FPS (43 ms latency), our method is approximately 1.9x faster than VideoMAE and 5.7x faster than frame-wise YOLOv8, demonstrating its suitability for real-time clinical monitoring.

3.2 Introduction

Monitoring vital signs in pediatric patients within Pediatric Intensive Care Units (PICUs) is essential due to their fragile health conditions. Non-contact approaches, such as remote photoplethysmography (rPPG) and respiratory monitoring, especially for conditions like Acute Respiratory Distress Syndrome (ARDS) (Poh *et al.*, 2010b; Khemani, Smith, Zimmerman, Erickson & Group, 2015), depend on accurate detection of anatomical regions, particularly the face and thoracoabdominal areas. Accurate and anatomically consistent localization is essential for reliable vital sign estimation in PICU environments. Remote photoplethysmography (rPPG) depends on stable face crops that preserve skin-only regions, as chrominance fluctuations are easily corrupted by background leakage or bounding box drift. Even minor spatial jitter or rotation error can distort the rPPG signal, degrading heart rate accuracy and introducing artifacts into the frequency spectrum. Similarly, thoracoabdominal detection must capture the cyclic expansion of the chest along the correct orientation axis to extract respiratory motion cues. Boxes that drift or encompass non-thoracic regions, such as blankets or bedrails, risk obscuring the subtle periodic deformations that encode respiration.

The PICU setting presents several challenges: patients are frequently obscured by medical devices, lighting conditions, patient orientations, and there is a lack of annotated data necessary for training effective models for our specific clinical setting. Notably, no existing public video datasets capture both facial and thoracoabdominal regions simultaneously in PICU environments, necessitating the creation of our own dataset. In addition to RGB inputs, depth information provides complementary geometric context that improves robustness to occlusions and illumination variability, both of which are frequent in PICU environments. Unlike RGB, depth is invariant to lighting changes and can help distinguish foreground anatomical structures from background clutter such as tubing, blankets, or bedrails. This is particularly valuable when visual cues are weak or partially obstructed. In such cases, depth enhances the stability of region tracking and supports more reliable detection of subtle motion patterns. However, consistent acquisition and integration of depth data in real-world clinical settings remain non-trivial, requiring careful sensor placement, calibration, and synchronization with RGB streams to ensure reliable performance across patient conditions and hardware setups.

Data scarcity remains one of the main challenges in the medical domain, primarily due to strict privacy constraints and ethical considerations. Collecting large-scale, manually annotated video datasets in a clinical environment like the PICU is exceptionally difficult and costly. Furthermore, there is a significant domain gap between general-purpose videos and clinical data; features learned from datasets such as Kinetics-400 often fail to generalize to the unique PICU setting (Chen, Ma & Zheng, 2019), limiting model performance across different healthcare settings (Guan & Liu, 2021; Yang, Soltan & Clifton, 2022b). Downstream tasks in the PICU are especially challenging due to the subtle motion patterns and unique anatomical features of pediatric patient, where standard face detectors often fail because of underdeveloped facial structures and frequent occlusions from medical equipment (Grooby *et al.*, 2023b; Dosso, Kyrollos, Greenwood, Harrold & Green, 2022). To address both data limitations and domain discrepancies, self-supervised learning (SSL) provides a promising solution. Pre-training on unlabeled hospital videos allows models to learn relevant spatiotemporal features directly from the target environment, reducing dependence on manual annotations (Xu *et al.*, 2019). Among SSL methods, masked autoencoders such as VideoMAE (Tong, Song, Wang & Wang, 2022) have demonstrated strong efficiency, offering robust representation learning for clinical settings where annotated data is limited.

While convolutional neural networks (CNNs) have significantly advanced static object detection and tracking, with strong performance on large-scale datasets such as COCO (Lin *et al.*, 2014) and PASCAL VOC (Everingham, Van Gool, Williams, Winn & Zisserman, 2010), these models are not designed to capture the temporal dynamics essential for physiological monitoring (Litjens *et al.*, 2017). Reliable signal monitoring depends on stable region-of-interest (ROI) tracking across time (Poh *et al.*, 2010a; Verkruysse *et al.*, 2008). Frame-based detectors, including recent YOLO variants, are limited in this context for two main reasons: first, they often suffer from temporal inconsistencies such as bounding box jitter, which introduces motion artifacts into the predicted signals (Bewley, Ge, Ott, Ramos & Upcroft, 2016; Zhang, Cheng, Zhu, Lin & Dai, 2018); second, their per-frame inference leads to high computational overhead, making them less suitable for real-time applications.

The video Vision Transformers (ViViT) model (Arnab *et al.*, 2021; Bertasius *et al.*, 2021; Fan *et al.*, 2021) rely on self-attention mechanisms to capture long-range dependencies and global context. However, they are computationally expensive and often struggle to generalize, especially when trained on limited datasets. Their lack of inherent inductive biases, such as spatial locality, makes it less effective in data-constrained environments. Moreover, its resource requirements increase quadratically when processing high-resolution inputs or integrating multimodal data like RGB and depth, which makes them difficult to deploy in real-world clinical settings. In contrast, 3D convolutional neural networks (3D-CNNs) (Tran, Bourdev, Fergus, Torresani & Paluri, 2015; Carreira & Zisserman, 2017) possess strong inductive biases, making them more suitable for limited-data scenarios. However, their localized kernel limit their ability to model long-range spatiotemporal dependencies, which are crucial for understanding complex clinical settings.

PICU monitoring systems must process multiple patient video streams continuously on shared, resource-limited workstations, often without dedicated accelerators or cloud offloading due to privacy and maintenance constraints. Reliable bedside use therefore requires low-jitter, real-time inference with a small memory footprint while coexisting with other clinical software. Under these conditions, models whose time and memory scale quadratically with the token length L (e.g., multi-head self-attention, $O(L^2)$) become impractical as resolution or temporal context grows. This motivates architectures with linear-time complexity ($O(L)$) and low VRAM that maintain stable latency on longer clips and higher resolutions.

Reliable face and thoracoabdominal localization is a prerequisite for contactless vital-sign monitoring in the PICU. The face region provides the skin pixels needed for rPPG, where minute chrominance fluctuations encode heart rate; any drift or background leakage rapidly degrades signal-to-noise and produces unstable frequency peaks. The thoracoabdominal region carries respiratory motion, so bounding boxes must remain temporally stable and orientation-aware to capture cyclic expansion/deflation without being contaminated by blankets, caregiver hands, or attached devices. Axis-aligned boxes are often insufficient under infant pose changes, bed tilt, and off-axis cameras; oriented bounding boxes (OBBs) yield tighter, rotation-consistent crops, improving both rPPG sampling using skin-only pixels and respiration estimation with motion along the anteroposterior axis. These constraints, coupled with frequent occlusions, specular lighting, and the need for real-time processing on clinical hardware, motivate a detector that is robust, temporally stable, and compute-efficient for face and thoracoabdominal ROIs, precisely the focus of our approach below.

To address these challenges, we introduce Divided Space–Time (DST) Mamba, an SSM-based detector for face and thoracoabdominal regions with oriented bounding boxes (OBBs). The model is built on Selective State Space Models (SSMs) (Gu & Dao, 2023) and runs in linear time $O(L)$ with respect to sequence length L, which is essential for real-time monitoring. Unlike VideoMamba, which processes space and time jointly, DST decouples them: a spatial stage followed by a temporal stage. This factorization reduces cross-axis interference, preserves temporal dynamics relevant to rPPG/respiration, and enables axis-specific optimization. It also allows independent MAE pre-training for spatial masked patches and temporal masked tubes objectives. To handle PICU conditions, like occlusions and low contrast, we use data-efficient masked-autoencoding pre-training and support multimodal input (RGB + depth) to improve perceptual robustness.

The primary contributions of this work are outlined as follows:

- Reliable ROI detection in PICU videos: we successfully detect the face and thoracoabdominal regions with oriented boxes despite occlusions, devices, motion, and limited labels in the PICU. This robustness is achieved via a Divided Space–Time Mamba design that preserves temporal dynamics while remaining compute-efficient.
- Data scarcity and domain gaps addressed: we mitigate scarce annotations and lab-to-PICU domain shift by employing self-supervised masked-autoencoder pre-training tailored to our clinical video distribution.
- Multimodal robustness under occlusion: we enhance detection accuracy and reduce angle drift in occlusion-heavy scenes by integrating RGB with depth and analyzing the accuracy–complexity trade-off for deployment.
- Clinical-grade efficiency and comparative gains: we achieve real-time throughput and low FLOPs while outperforming strong frame-wise and video baselines on accuracy and temporal stability through a factorized spatial-to-temporal SSM pipeline and targeted ablations.

3.3 Related Works

3.3.1 Object Detection in Videos

Deep learning techniques have significantly advanced object detection. Initially, two-stage detectors like R-CNN(Girshick, Donahue, Darrell & Malik, 2014), Fast R-CNN(Girshick, 2015) and Faster R-CNN (Ren, He, Girshick & Sun, 2015) utilized region proposals to achieve high accuracy. Single-stage detectors, including the YOLO family of models (Redmon, Divvala, Girshick & Farhadi, 2016; Redmon & Farhadi, 2017, 2018) and the Single Shot Multibox Detector (SSD)(Liu *et al.*, 2016), emerged as faster alternatives by predicting bounding boxes and class probabilities in a single pass. More recently, transformer-based approaches such as DETR (Carion *et al.*, 2020) and Deformable DETR (Zhu *et al.*, 2020) have been proposed, leveraging global context through Multi-Head Self-Attention to enable end-to-end object detection.

However, these methods process frames independently, relying on separate tracking modules (e.g., YOLO + DeepSORT (Bewley *et al.*, 2016)) for temporal coherence, which increases latency and introduces jitter or artifacts that degrade vital sign signals in critical care (Álvarez Casado, Nguyen, Silvén & Bordallo López, 2023a; Wang, Shan, Liu, Zhou & Shu, 2025). This gap highlights the importance of using factorized spatiotemporal modeling to maintain temporal stability and orientation-aware detection under real-time constraints. By avoiding reliance on external trackers and ensuring linear complexity, such designs support consistent performance in PICU monitoring scenarios.

3.3.2 Face Detection in Complex Environments: NICU/PICU

Face detection in NICU/PICU settings has evolved from traditional hand-crafted methods like Haar cascades with AdaBoost (Bradski, 2000) to deep learning models such as MTCNN (Zhang, Zhang, Li & Qiao, 2016), RetinaFace (Deng, Guo, Ververas, Kotsia & Zafeiriou, 2020), BlazeFace (Bazarevsky, Kartynnik, Vakunov, Raveendran & Grundmann, 2019), and YOLO5Face (Qi, Tan, Yao & Liu, 2022), which offer robustness to occlusions and real-time performance.

Standard detectors struggle with neonatal morphology, medical occlusions, and cluttered backgrounds, as shown in NICU-specific adaptations such as NICU-Face (YOLOv5-based) (Dosso *et al.*, 2022), Hausmann's model (Hausmann, Salekin, Zamzmi, Goldgof & Sun, 2022) , and Grooby's YOLOv7 (Grooby *et al.*, 2023a). Integrated approaches for vital sign estimation, including Huang et al. (Huang *et al.*, 2021b) for heart rate and Kyrollos et al. (Kyrollos, Tanner, Greenwood, Harrold & Green, 2021) for respiration, still lack temporal consistency, often resulting in unstable region tracking and noisy physiological signals. In contrast, our approach leverages self-supervised pre-training on domain-specific video data and incorporates multimodal RGB-D input to improve robustness to occlusions and ensure stable tracking of anatomical regions for continuous monitoring.

3.3.3 Thoracoabdominal Detection and Respiratory ROI Tracking

Thoracoabdominal detection for respiratory monitoring often relies on depth sensors and classical approaches, including infrared imaging for torso motion (Eastwood-Sutherland, Lim, Gale, Wheeler & Dargaville, 2023), time-of-flight cameras with anatomical landmarks (Rong & Bliss, 2023), and segmentation-based techniques such as normalized cuts (Kaur *et al.*, 2017) or probability masks (Nagy *et al.*, 2021). Rehouma et al. (Rehouma, Noumeir, Bouachir, Jouvet & Essouri, 2018) reconstructed a 3D thoracoabdominal surface using dual Kinect v2 sensors to capture respiratory patterns. Simpler pixel-tracking methods (Massaroni, Lo Presti, Formica, Silvestri & Schena, 2019; L'Her, Nazir, Pateau & Visvikis, 2022) and static deep learning models (Ronneberger, Fischer & Brox, 2015; Shen, Zhou, Yang, Yang & Tian, 2015) have also been proposed, although they lack the ability to capture temporal dynamics.

Frame-based approaches fail to preserve periodic respiratory motion under occlusion, emphasizing the need for spatiotemporal continuity. The decoupled temporal design employed in our method captures subtle inter-frame motion patterns, while the use of oriented bounding boxes enables rotation-consistent localization in occlusion-heavy PICU environments.

3.3.4 Video Understanding Models

To address the need for temporal continuity, models that process video data have been developed. Three-dimensional convolutional neural networks (3D-CNNs) (Tran *et al.*, 2015; Carreira & Zisserman, 2017) extend standard CNNs into the temporal dimension by convolving across successive frames. While 3D-CNNs effectively learn short-term motion features, their fixed temporal receptive field limits their ability to capture long-range dependencies, such as full breathing cycles or prolonged occlusions. Increasing their depth or temporal window significantly raises computational costs, making them impractical for long PICU video sequences (Li, Zhang, Liu, Lei & Li, 2023a; Fayyaz *et al.*, 2021; Hedegaard & Iosifidis, 2022).

More recent transformer-based video models, such as ViViT (Arnab *et al.*, 2021) and TimeSformer (Bertasius *et al.*, 2021), apply self-attention to sequences of frame patches, effectively modeling global spatiotemporal relationships. These approaches achieve strong performance on action recognition benchmarks by capturing interactions across entire clips. However, their self-attention mechanism has quadratic complexity with respect to the number of tokens (spatial patches × temporal frames), resulting in high memory and computational demands. For example, processing a 30-s PICU video at clinically meaningful resolution would involve attending over a high-dimensional token sequence, making such models impractical for real-time deployment without specialized hardware.

3.3.5 Self-Supervised Video Representation Learning

In data-limited environments, self-supervised learning (SSL) offers an efficient strategy for pre-training models on unlabeled videos by constructing surrogate tasks. This is particularly relevant in medical contexts, where data collection is constrained by privacy concerns and manual annotation is costly (Xu *et al.*, 2019). By learning directly from the data, SSL enables models to acquire meaningful representations that can be transferred to downstream tasks such as detection or segmentation. A common SSL approach for video is contrastive learning, where models are trained to map different augmentations of the same video clip to similar embeddings, while pushing apart embeddings from different clips. Momentum Contrast (MoCo) (He, Fan, Wu, Xie & Girshick, 2020) and SimCLR (Chen, Kornblith, Norouzi & Hinton, 2020b) are prominent examples using instance discrimination objectives. More recent variants such as BYOL (Grill *et al.*, 2020) and DINO (Caron *et al.*, 2021) eliminate the need for explicit negative samples by employing teacher–student architectures to learn invariant features from augmented video data.

Another class of self-supervised learning (SSL) methods is generative or reconstruction-based. These approaches involve masking or removing parts of the input and training the model to predict the missing content, thereby encouraging it to learn contextual and semantic structures. Masked image modeling (MIM) techniques, inspired by BERT in natural language processing (NLP) (Devlin, Chang, Lee & Toutanova, 2019), have shown strong performance in both image and video domains. For example, iGPT (Chen *et al.*, 2020a) and BEiT (Bao, Dong, Piao & Wei, 2021) demonstrated the effectiveness of tokenizing images and learning through masked token prediction. In particular, Masked Autoencoders (MAE) (He *et al.*, 2022) showed that a vision transformer can be pre-trained efficiently by encoding only a small subset of visible image patches and training a lightweight decoder to reconstruct the missing ones. VideoMAE (Tong *et al.*, 2022) extended this concept to video by leveraging the high redundancy between frames. It employs an extremely high masking ratio (90–95%) using a tube masking strategy, which masks consistent spatial regions across consecutive frames, enabling efficient learning of spatiotemporal representations. A common challenge with reconstruction-based self-supervised learning (SSL) is that optimizing for low-level pixel accuracy may not produce representations that capture high-level semantic features. Recent advancements, such as Unmasked Teacher (UMT) (Li *et al.*, 2023b), address this limitation by incorporating a teacher network that identifies informative tokens and provides softer reconstruction targets, thereby guiding masked autoencoders toward learning more semantic representations.

Nonetheless, a key advantage of the MAE approach in our context is that it enables pre-training directly on our collected PICU video data, as well as on additional video sequences created from real clinical images captured in PICU/NICU settings, without requiring any labels. This allows the model to learn representations adapted to the hospital environment, such as the appearance of neonatal skin under PICU lighting or the typical motion patterns of breathing infants, effectively bridging the domain gap encountered when using models pre-trained on general-purpose video datasets such as Kinetics-400.

3.3.6 State Space Models (SSMs)

Transformers have become the dominant architecture for sequence modeling in both natural language processing (NLP) and computer vision due to their capacity for global attention (Vaswani *et al.*, 2017). However, their $O(n^2)$ complexity with respect to sequence length makes them less tractable for very long sequences or high-resolution video. State Space Models (SSMs) offer an alternative sequence modeling paradigm with $O(n)$ complexity, based on simulating linear dynamical systems. The Structured State Space Sequence Model (S4)(Gu, Goel & Ré, 2021) introduced a parameterization that enables learning long-range dependencies via a diagonal-plus-low-rank representation of the state transition matrix, achieving strong performance on long-sequence tasks while maintaining linear time complexity. Subsequent refinements, including S5(Smith, Warrington & Linderman, 2022), H3 (Fu *et al.*, 2022), and GSS (Mehta, Gupta, Cutkosky & Neyshabur, 2022), further improved the stability and efficiency of SSM-based sequence layers.

Mamba (Gu & Dao, 2023), based on the State Space Model (SSM), introduced a data-dependent SSM layer that enables efficient processing of long sequences while maintaining computational efficiency. It incorporates Selective State Spaces by making the state transition matrices input-dependent, allowing the model to selectively process information based on the current input. In addition, it employs hardware-aware parallelism to optimize long-sequence processing by avoiding unnecessary memory access through selective scans and kernel fusion. This design minimizes latency, maximizes throughput on modern GPUs, and achieves true linear-time scaling. Mamba outperforms Transformer architectures on large-scale real-world datasets and scales linearly with sequence length. Recently, several Mamba-based approaches have leveraged the strengths of SSMs to efficiently model long sequences (Pióro *et al.*, 2024; Zhao *et al.*, 2025).

The Mamba architecture has been extended to computer vision tasks through several adaptations. Vision Mamba (ViM)(Zhu *et al.*, 2024) generalizes Mamba from 1D to 2D sequences by employing bidirectional scans, processing all tokens in both forward and backward directions to enhance spatial representations. VMamba(Liu *et al.*, 2024c) introduces a different scanning strategy using 2D Selective Scan (SS2D), which processes tokens in four directions to enrich contextual information. EfficientVMamba (Pei, Huang & Xu, 2025) proposes a lightweight model that applies atrous sampling on feature map patches to reduce computational complexity. To improve local representation, LocalMamba (Huang *et al.*, 2024) divides the image into groups and scans each group's window independently. It also incorporates spatial and channel attention modules to filter out redundant information and retain only the most relevant features. The strong performance of Mamba-based backbones across diverse vision tasks has spurred the development of specialized models tailored to specific applications (Xing, Ye, Yang, Liu & Zhu, 2024; Wang, Tsepa, Ma & Wang, 2024b; Ruan, Li & Xiang, 2024; Liu *et al.*, 2024a; Behrouz & Hashemi, 2024).

3D convolutional neural networks (3D CNNs) are computationally expensive and memory-intensive, while video transformers suffer from poor scalability due to the quadratic complexity of self-attention with respect to input length. Both approaches tend to be slow, resource-demanding, and require substantial amounts of training data. To address these limitations, Mamba has recently been extended to video understanding tasks, offering a more efficient alternative to 3D CNNs and video transformers. VideoMamba (Li *et al.*, 2024) is designed to maintain linear complexity for long-range video modeling. It begins by dividing the input video into non-overlapping spatiotemporal patches using a 3D CNN, followed by the addition of learnable spatial and temporal positional embeddings. Spatial tokens are arranged according to their locations and stacked sequentially across frames. Leveraging Mamba's linear-time Selective State Space mechanism, VideoMamba can efficiently process long, high-resolution video sequences.

However, due to the phenomenon of historical decay, where earlier tokens have limited influence on later outputs; VideoMambaPro (Lu, Salah & Poppe, 2024) improves upon the original model by introducing masked backward computation in the bidirectional Mamba process and residual connections within the Mamba transition matrices.

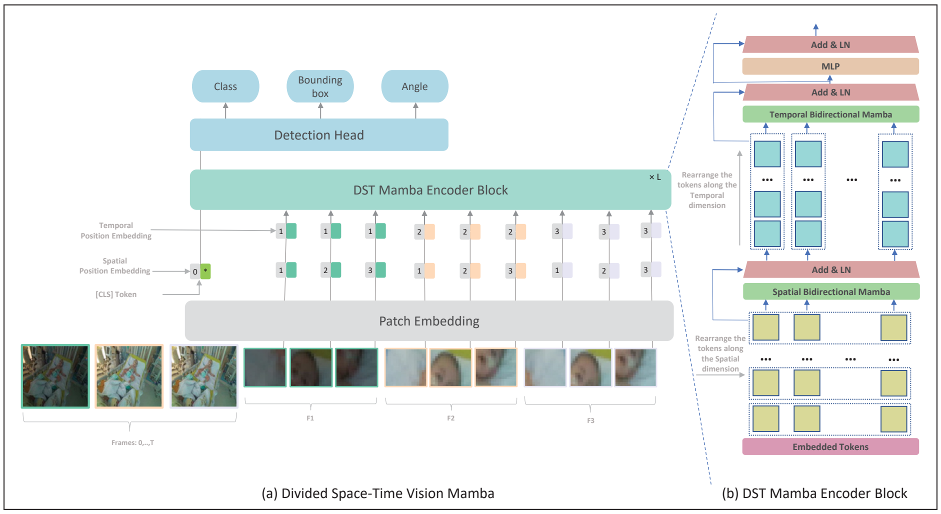
In contrast, we propose a Divided Space–Time Mamba architecture that explicitly decouples spatial and temporal sequence modeling. Inspired by the factorized space-time attention in TimeSformer (Bertasius *et al.*, 2021), our model processes spatial and temporal information in separate stages: spatial Mamba layers first operate within each frame to preserve high-resolution spatial details, followed by temporal Mamba layers that model dependencies across frames using the spatially encoded features. The factorized space–time design enables the model to develop specialized representations along each axis, preserving high-resolution anatomical detail through spatial encoding while capturing temporal dynamics. By decoupling spatial and temporal processing, the architecture avoids the representational and computational trade-offs inherent in joint spatiotemporal models. This separation prevents subtle, time-sensitive signals from being overwhelmed by dominant spatial features, a limitation often observed in unified attention-based approaches.

3.4 Proposed Method

3.4.1 Overview

In this work, we propose a Mamba-based approach for medical video detection that reliably localizes the face and thoracoabdominal regions in PICU videos. The backbone factorizes spatiotemporal modeling by stacking Divided Space–Time (DST) Mamba blocks, which first consolidate spatial structure before modeling temporal dynamics. This sequential design allows the encoder to capture both coarse scene layouts and fine-grained motion cues, all while preserving temporal signals with linear-time complexity. On these features, a lightweight Mamba-based detection head predicts oriented bounding boxes (OBBs), yielding rotation-consistent localization under pose changes, bed tilt, and off-axis cameras, as shown in Figure 3.1.

To address data scarcity and domain shift, we investigate two SSL pre-training techniques: Masked Autoencoders (MAE) and Unmasked Teacher (UMT). MAE reconstructs masked patches to learn robust local appearance priors, whereas UMT distills higher-level spatiotemporal structure from a teacher without masking. We quantify the contribution of SSL and depth by comparing VideoMamba and our DST-Mamba with and without pretraining and with RGB versus RGB-D input. All models are first pretrained on a combined set of CHU Sainte-Justine PICU clips and publicly available pediatric data, then fine-tuned on the PICU dataset for face and thoracoabdominal OBB detection. This comparative design targets the core clinical video challenges of data scarcity, heavy occlusions, and rapid domain shift, while favoring deployment through linear-time inference.


Class
Bounding box
Angle
Detection Head
DST Mamba Encoder Block
× L
Temporal Position Embedding
Spatial Position Embedding
[CLS] Token
0
*
1
2
3
Patch Embedding
F1
F2
F3
Frames: 0,..,T
Add & LN
MLP
Add & LN
Temporal Bidirectional Mamba
Rearrange the tokens along the Temporal dimension
Add & LN
Spatial Bidirectional Mamba
Rearrange the tokens along the Spatial dimension
Embedded Tokens

(a) Divided Space-Time Vision Mamba

(b) DST Mamba Encoder Block

Figure 3.1 Overview of our framework: (**a**) Divided Space–Time Video Mamba integrates patch embedding, positional embeddings, Mamba encoder layers, and a detection head for class labels, bounding boxes, and angles. (**b**) The DST Mamba Block. Input tokens are first processed by a spatial Mamba layer that operates within each individual frame. The output is then reshaped and processed by a temporal B-Mamba layer across all frames. * denotes the CLS token position

3.4.2 Preliminary Explanation: State Space Models

State Space Models (SSMs) map 1-D function or sequence $x(t) \in \mathbb{R} \mapsto y(t)$ using a hidden state $h(t) \in \mathbb{R}^N$. This system is described as linear ordinary differential equations (ODEs), employing matrices $\mathbf{A} \in \mathbb{R}^{N\times N}$ to define how the hidden state evolves and $\mathbf{B} \in \mathbb{R}^{N\times 1}$ and $\mathbf{C} \in \mathbb{R}^{1\times N}$ for the projection of the input and the hidden state to the output:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t). \end{aligned} \tag{3.1}$$

S4 (Gu *et al.*, 2021) and Mamba (Gu & Dao, 2023) integrate a timescale parameter $\mathbf{\Delta}$ to discretize the continuous system and convert the continuous parameters $\mathbf{A}, \mathbf{B}$ to discrete parameters $\bar{\mathbf{A}}, \bar{\mathbf{B}}$. The transformation defined as follows:

$$\bar{\mathbf{A}} = \exp(\mathbf{\Delta A}), \quad \bar{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I}) \cdot \mathbf{\Delta B}. \tag{3.2}$$

After the discretization of $\bar{\mathbf{A}}, \bar{\mathbf{B}}$, the (3.1) is transformed into:

$$\begin{aligned} h_t &= \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \\ y_t &= \mathbf{C}h_t. \end{aligned} \tag{3.3}$$

A global convolution employed to compute the model output:

$$\begin{aligned} \overline{\mathbf{K}} &= (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{AB}}, \dots, \mathbf{C}\overline{\mathbf{A}}^{\mathbf{M-1}}\overline{\mathbf{B}}), \\ y &= x * \bar{\mathbf{K}}. \end{aligned} \tag{3.4}$$

Here M represents the length of the input sequence x, and $\overline{\mathbf{K}} \in \mathbb{R}^M$ is a structured convolutional kernel.

3.4.3 Divided Space–Time Video Mamba

3.4.3.1 Baseline: Joint Spatiotemporal Processing

We first implemented a baseline following VideoMamba (Li *et al.*, 2024) which processes spatial and temporal information jointly through a unified bidirectional scanning mechanism, Fig. 3.2. VideoMamba extends the Mamba state space model to video understanding by treating the entire video as a single sequence of spatiotemporal tokens. Given an input video $\mathbf{X} \in \mathbb{R}^{3\times T\times H\times W}$, where T is the number of frames and $H \times W$ are spatial dimensions, VideoMamba first applies a 3D convolutional patch embedding to obtain N spatiotemporal patches $\mathbf{X}_p \in \mathbb{R}^{N\times D}$; where $N = \frac{T\cdot H\cdot W}{P^2}$ for P is the patch size , and D is the embedding dimension. Each token represents a local spatiotemporal cube containing information from multiple consecutive frames. VideoMamba (Li *et al.*, 2024) applies joint scanning strategy. All N tokens are arranged in a single sequence according to a spatial-first ordering:

$$\mathbf{S}_{\text{joint}} = [\mathbf{x}_{1,1}, \mathbf{x}_{2,1}, \ldots, \mathbf{x}_{HW/P^2,1}, \mathbf{x}_{1,2}, \ldots, \mathbf{x}_{HW/P^2,T}] \tag{3.5}$$

This sequence is then processed by bidirectional Mamba blocks:

$$\mathbf{Y}_{\text{forward}} = \text{SSM}_{\text{forward}}(\mathbf{S}_{\text{joint}}; \mathbf{A}, \mathbf{B}, \mathbf{C}, \Delta) \tag{3.6}$$

$$\mathbf{Y}_{\text{backward}} = \text{SSM}_{\text{backward}}(\mathbf{S}_{\text{joint}}; \mathbf{A}, \mathbf{B}, \mathbf{C}, \Delta) \tag{3.7}$$

$$\mathbf{Y}_{\text{joint}} = \mathbf{Y}_{\text{forward}} + \mathbf{Y}_{\text{backward}} \tag{3.8}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$ are the state space parameters, and Δ is the time-scale parameter. The bidirectional scan enables each token to aggregate context from both past and future tokens in the sequence. All the patches are processed then using L Bidirectional Mamba blocks where a spatial-first bidirectional scan is applied.The joint approach may not allow for fine-tuning the balance between spatial and temporal processing. By processing spatial and temporal information jointly, the model might not develop as specialized features for each dimension.

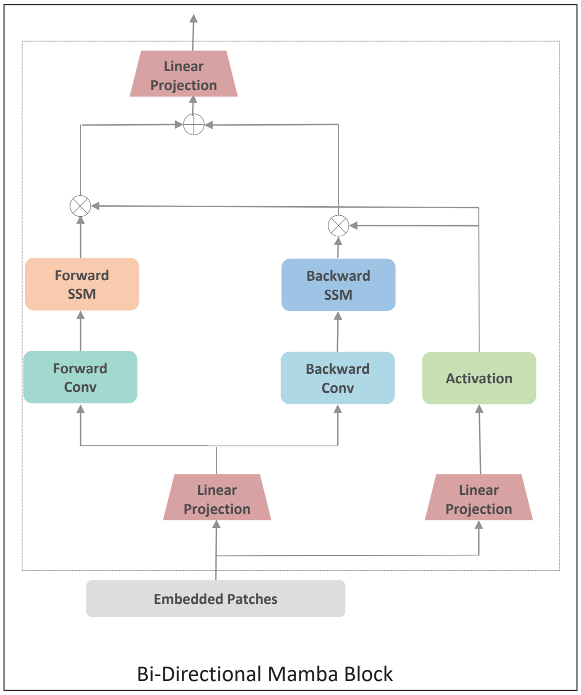


Figure 3.2 Detailed architecture of the Bidirectional Mamba block used for joint spatiotemporal processing. The input sequence of Embedded Patches is processed through two parallel streams to capture dependencies in both forward and backward directions

3.4.3.2 Divided Space–Time Processing

While TimeSformer (Bertasius *et al.*, 2021) employs a Divided Space–Time Multi-Head Self-Attention (MHSA), its quadratic complexity with respect to token count poses challenges for long video sequences, where token numbers grow linearly with input frames. To address this, we propose a Divided Space–Time Mamba block that models intra- and inter-frame long-range dependencies efficiently, resolving scalability issues without sacrificing performance. We refined this approach by introducing a modified Vision Mamba architecture based on a Divided

Space–Time Mamba model, as illustrated in Fig. 3.1. By separating Vision Mamba into spatial and temporal modules, the architecture leverages specialized learning for each dimension: the spatial module captures fine-grained details within individual frames, while the temporal module tracks movement and event progression over time. This division is particularly effective for medical video detection, enabling the model to learn dynamic appearance and motion cues more efficiently.

Each frame in the input clip is divided into non-overlapping patches of size $P \times P$. This ensures that the N patches cover the entire frame, with N defined as $N = HW/P^2$. Each token is represented by $\mathbf{x}_{(p,t)} \in \mathbb{R}^{3P^2}$, where p and t are spatial locations and frame index. The sequence of tokens is initially arranged in $X \in \mathbb{R}^{N \times T \times D}$, where N represents the patch position within each frame, T indexes time and D is the embedding size of each token.

The encoder blocks process temporal and spatial dimensions separately, one after the other. Each block l, we first use $X^{space} \in \mathbb{R}^{(B \times T) \times N \times D}$ to fix the temporal dimension. Then, we perform a bidirectional scan across all frames to capture spatial dependencies:

$$y^{space}(t) = \mathrm{SSM}_{spatial}(X^{space}(t)). \tag{3.9}$$

The output of temporal B-Mamba scan is then feed forward to compute temporal B-Mamba encoder where all tokens are grouped based on frames $X^{time} \in \mathbb{R}^{(B \times N) \times T \times D}$. Across both time and space dimensions, separate parameters are learned: $\mathbf{A}^{time}, \mathbf{B}^{time}, \mathbf{C}^{time}$ for the temporal component and $\mathbf{A}^{space}, \mathbf{B}^{space}, \mathbf{C}^{space}$ for the spatial component.

3.4.4 Pretraining Approaches

To address the scarcity of labeled data in clinical video settings, we investigated two distinct self-supervised learning (SSL) strategies for initializing our Divided Space–Time Mamba model: a Teacher–Student approach based on semantic distillation, and a fully self-supervised reconstruction strategy using masked autoencoding. The embedded tokens are passed through the encoder and decoder parts, respectively. The encoder part consists of L stacked Mamba blocks and aims to extract meaningful latent representations by processing only masked input sequences. These representations capture the context and structure of the visible data while learning to predict missing tokens. In the pretraining stage, the learnable special token is removed.

Teacher-Student SSL: We first attempted a teacher-student SSL approach using CLIP (Radford *et al.*, 2021) as the teacher model providing semantic guidance. Inspired by previous works (Li *et al.*, 2023b, 2024), the decoder part aligns unmasked tokens directly with a linear projection to the teacher model. For masking strategy, we employ a frame-by-frame semantic masking approach, assigning higher probabilities to tokens that carry crucial clues (Li *et al.*, 2023b; Hou, Sun, Chen, Xie & Kung, 2022). In the PICU setting, this strategy resulted in unstable convergence and poor generalization, which is caused by: the semantic gap between the teacher and the target domain, and the mismatch between the pretraining objectives. CLIP was trained on generic web images and captions that emphasize everyday objects and scenes, while our clinical data involve subtle and domain-specific patterns such as neonatal anatomy, occlusions, and the presence of medical equipment. The teacher often highlighted irrelevant elements like monitors or cables instead of the anatomical regions required for detection. Additionally, CLIP's training objective focuses on global image–text alignment, whereas our model requires precise spatial localization and temporal consistency. This misalignment likely led to conflicting gradients during training, especially under high masking ratios and in visually degraded frames.

Masked Autoencoders SSL: UMT relies heavily on the semantic guidance from the teacher model. If this model is pre-trained on general images that are vastly different from PICU environments, the guidance might be less relevant or even misleading. Video Masked Autoencoders's self-supervised approach allows it to learn directly from the target domain (PICU videos) without relying on potentially mismatched external knowledge. The decoder part consists of stacked B-Mamba blocks with a final output projection to reconstruct the masked video patches. VideoMAE's masking strategy and reconstruction objective enable it to capture domain-specific features and patterns present in PICU data, even if they're very different from general image datasets. In the Divided Space–Time Mamba model, and to keep the structure, masked tokens are replaced with learnable parameters. These learnable embeddings act as placeholders for the missing information and are processed alongside the unmasked tokens. The learnable parameters allow gradients to flow through the masked positions, which can improve model optimization.

3.4.5 Depth Information Integration

To enhance the ability of the model to learn subtle anatomical features and motion cues within the complex PICU setting. We augment our Divided Space–Time Mamba architecture by introducing depth maps as an additional input channel alongside RGB frames. For each input video frame $\mathbf{X}_t \in \mathbb{R}^{3\times H\times W}$, we incorporate a corresponding depth map $\mathbf{D}t \in \mathbb{R}^{1\times H\times W}$, resulting in a four-channel input $\mathbf{X}_t^d \in \mathbb{R}^{4\times H\times W}$. Each token is represented by $\mathbf{x}_{(p,t)} \in \mathbb{R}^{4P^2}$.

Depth information is fused through early channel-level concatenation prior to tokenization, ensuring pixel-wise alignment between RGB and depth modalities. The fused 4-channel frames are processed by a shared patch-embedding layer and the same Divided Space–Time Mamba encoder, without the use of a separate fusion or attention branch.

During the pretraining phase, we apply the same masking strategies to both RGB and depth channels, encouraging the model to learn the relationships between appearance and geometric features. Our experiments demonstrate that the incorporation of depth information leads to improved performance in both pretraining and downstream tasks, particularly in scenarios requiring precise spatial understanding of the PICU environment.

3.4.6 Fine-Tuning

The pretrained model undergoes fine-tuning on the PICU dataset for face and thoracoabdominal detection. During fine-tuning, the decoder block is replaced with a lightweight detection head comprising three projection layers for classification, bounding box regression, and orientation angle prediction.

The detection task in PICU environments necessitates a multi-component loss function to address distinct challenges inherent to clinical video analysis. The total loss integrates four components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{angle}} + \alpha \cdot \mathcal{L}_{\text{bbox}} + \beta \cdot \mathcal{L}_{\text{IoU}} \tag{3.10}$$

where α and β balance the contribution of each component based on their typical value ranges during training.

Classification Loss ($\mathcal{L}_{\text{cls}}$): Binary Cross-Entropy is employed for object presence prediction. This choice addresses the non-mutually exclusive nature of face and thorax detection in PICU frames, where medical equipment frequently occludes one region while leaving the other visible.

Angle Loss ($\mathcal{L}_{\text{angle}}$): Orientation prediction utilizes Cross-Entropy loss with Circular Spatial Layout (CSL), discretizing the 180° rotation space into 180 bins. The CSL formulation addresses the periodicity of angular measurements, where standard Cross-Entropy would incorrectly treat adjacent angles (e.g., 179° and 1°) as maximally different.

Bounding Box Regression ($\mathcal{L}_{\text{bbox}}$): L1 loss optimizes the coordinate predictions for bounding box parameters (x, y, w, h). We selected L1 over L2 loss due to its robustness to annotation outliers present in clinical data.

Oriented IoU Loss ($\mathcal{L}_{\text{IoU}}$): The rotated Intersection over Union loss directly optimizes the spatial overlap between predicted and ground truth oriented bounding boxes:

$$\mathcal{L}_{\text{IoU}} = 1 - \frac{\text{Area}(B_{\text{pred}} \cap B_{\text{gt}})}{\text{Area}(B_{\text{pred}} \cup B_{\text{gt}})} \tag{3.11}$$

This component ensures spatial alignment beyond coordinate accuracy, particularly crucial for oriented boxes where axis-aligned IoU would penalize correctly oriented predictions. In cases involving patient rotation, bed tilt, or oblique camera angles, a predicted box may be geometrically correct yet misjudged by axis-aligned IoU metrics due to misalignment with image axes. The rotated IoU (rIoU) metric accounts for both position and orientation, providing a more accurate measure of overlap under rotation. This is particularly important for respiratory motion analysis, where chest orientation must be tracked precisely, and penalizing correctly rotated predictions would compromise model evaluation and training.

3.5 Experimental Protocol

3.5.1 Data Acquisition

To evaluate our approach, we collected two datasets. The first was gathered at the PICU of CHU-Sainte-Justine Hospital (CHU-SJ). To our knowledge, this represents the only available video dataset that captures both facial and thoracoabdominal regions in a PICU setting, as existing datasets typically focus on either face detection or respiratory monitoring in isolation. Due to privacy and ethical constraints, acquiring large-scale annotated PICU video datasets is challenging. Our method overcomes this by generating video-like data from publicly available images, enabling effective pre-training with limited real video data.

3.5.1.1 CHU-SJ Videos Collection

At CHU-Sainte-Justine Hospital's PICU, videos are collected using a Microsoft Azure RGB-D sensor color camera with 30 FPS (ultra-HD 12-megapixel RGB camera). Approximately 485 different patients admitted to the PICU of CHU-SJ were recorded for 30 seconds each (Boivin *et al.*, 2023). Patients, especially infants and young children, frequently moved or shifted positions, causing their faces and thoraxes to move out of the camera's field of view. Variability in lighting, including low light during nighttime or shadows from medical equipment, significantly affected video quality. The presence of medical devices such as ventilator tubes, masks, or monitoring leads often obscured key regions, complicating the detection process. To quantify these environmental challenges, Figure 3.3 provides a statistical breakdown of common occlusion sources. As shown, medical necessities like oxygen masks (6.0%) and patient coverings such as cloths (5.8%) and hats (4.4%) are the most significant contributors, underscoring the need for an occlusion-robust detection model.

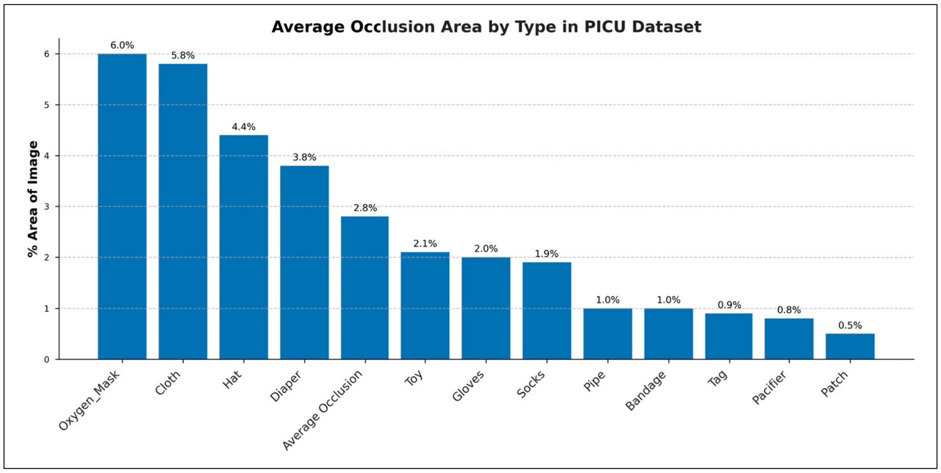


Figure 3.3 Average image area occluded by common sources in the CHU-SJ PICU dataset. Medical equipment (e.g., oxygen masks, 6.0%) and patient coverings (e.g., cloths, 5.8%) are the dominant occlusion types

Faces and thoracoabdominal regions within each video frame were manually annotated using oriented bounding boxes. For face detection, the oriented bounding box was drawn around the face, from the forehead to the chin and from ear to ear. For thoracoabdominal detection, the oriented bounding box covered the area from the upper chest to the diaphragm, including the region of the thorax and abdomen. The dataset comprises individuals with varied attributes, such as skin color, ethnicity, and an age range spanning from 0 to 18 years. This diversity ensures that our dataset captures a wide spectrum of the population across skin tones, genders, and age groups. The institutional ethics committee of Sainte-Justine Hospital approved the study and database construction (protocol code 2016-1242) on March 31, 2016. Prior to video recording, parental consent was obtained by a research assistant trained in human ethics.

3.5.1.2 Public Data Collection

To address the scarcity of labeled data in medical settings, we constructed a new dataset derived from publicly available images with hospital settings, with a specific focus on neonatal, infant, pediatric, and intensive care units. Approximately 5,000 images were collected using these keywords and transformed into video sequences using a domain-specific sequential data augmentation strategy. To simulate the temporal and spatial dynamics of real videos, each image was duplicated into multiple frames, creating the illusion of continuity. Temporal variations, such as adjustments in brightness, contrast, and color, were applied progressively across frames to replicate real-world changes in lighting conditions and image intensity over time. Spatial variations were carefully introduced to enrich the dataset with dynamic visual effects. These included random shifts to simulate minor positional changes, random rotations to emulate natural camera movements, and the addition of subtle noise for enhanced texture realism. Collectively, these augmentations resulted in over 15k video clips, providing a rich resource for pre-training models. This approach addresses the challenge of limited annotated video data while maintaining high relevance to medical environments, thereby supporting feature learning for our specific environment.

3.5.1.3 Pre-Training Stage

We first conduct experiments on the self-supervised pre-training approaches using both datasets simultaneously. Each 30-second patient video is divided into video clips, sampled with a temporal stride of 4 to address the temporal redundancy often present in consecutive frames. Each video clip consists of 16 frames of size 224×224. For both SSL pre-training methods, we follow most of the hyperparameter settings described in (Tong *et al.*, 2022; Li *et al.*, 2023b), but we use a masking ratio of 80% for both of them. The models are pre-trained on a both datasets of $50k$ video clips using the AdamW optimizer over 2500 epochs. The training process employs a base learning rate of 1.5×10^{-4}, weight decay of 0.05, and warmup period of 40 epochs. We adjust the base learning rate in direct proportion to the total batch size, using the formula lr = base learning rate $\times$ batch size$/256$. The encoder architecture is set to a depth of 12 with an embedding dimension of 768, while the decoder has an embedding dimension of 384. To encourage the student model to learn high-level representations in Teacher-Student SSL, we use Mean Squared Error (MSE) to align unmasked tokens from the student with those from the teacher model. For the reconstruction task in Masked Autoencoders SSL, MSE is applied to measure the difference between the normalized masked pixels and their reconstructed counterparts.

3.5.1.4 Fine-Tuning Stage

The pre-trained models are fine-tuned on downstream tasks, specifically face and thoracoabdominal detection in the PICU. This evaluation assesses their effectiveness in learning from small, specialized datasets while demonstrating data efficiency, how effectively the model can train on a dataset with a complex setup. We evaluate the model's data efficiency and transfer learning capabilities through three training scenarios: (1) training from scratch, (2) pre-training on Kinetics-400 followed by fine-tuning on the PICU dataset, and (3) using our VideoMAE pre-trained weights based on Vision Transformer. The fine-tuning architecture consists of a Mamba Encoder backbone with specialized prediction heads for classification, bounding box regression, and rotation angle prediction with 180 categories for each detected object. In order to capture representative frames across the entire video, we employed a uniform sampling strategy with 16 segments per video. For CHU-SJ dataset, we generated 7216 samples, with the first 373 patients used for training and the remaining 112 subjects for testing. We report the results based on the remaining patients. Given the limited dataset size, we additionally performed 5-fold patient-wise cross-validation to verify model stability. Patients were randomly divided into 5 folds ensuring no patient appears in multiple folds. Each fold maintained approximately the same age distribution and occlusion severity. To assess the influence of frame resolution on model performance, we experiment with three resolutions: 224×224 and 640×640. The training was carried out using the Rectified Adam (RAdam) optimizer, with a learning rate of 1×10^{-3}. Training was conducted for 100 epochs, with a batch size of 32. All experiments were implemented in PyTorch version 2.0.1, and the network was trained on a single NVIDIA Tesla V100S-PCIE-32GB GPU.

3.5.2 Evaluation Metrics

3.5.2.1 Detection Metrics

We evaluate detection performance using mean Average Precision (mAP), following standard protocols adapted for oriented bounding boxes. This includes mAP at specific IoU thresholds of 0.50, 0.60, and 0.75, as well as the comprehensive mAP50-95 metric, which averages performance across IoU thresholds from 0.50 to 0.95. The rotated IoU (rIoU) is computed by determining the exact intersection area of the two convex polygons that define the boxes:

$$\text{rIoU} = \frac{\text{Area}(B_{pred} \cap B_{gt})}{\text{Area}(B_{pred} \cup B_{gt})}$$

where B_{pred} and B_{gt} are oriented rectangles parameterized by (x_c, y_c, w, h, θ).

3.5.2.2 Temporal Consistency Metrics

Temporal IoU measures detection stability across consecutive frames, quantifying tracking smoothness by averaging the IoU of an object's bounding box between adjacent frames. A higher value indicates less jitter.

$$\text{Temporal IoU} = \frac{1}{T-1} \sum_{t=1}^{T-1} \text{IoU}(B_t, B_{t+1})$$

where B_t represents the detected bounding box at frame t.

3.5.2.3 Angle Evaluation

Oriented bounding box angles are predicted using Circular Spatial Layout (CSL), which frames the task as a classification problem. We discretize the 180° orientation space into 180 classes, resulting in a 1° resolution. This approach, trained with a Cross-Entropy-based loss, naturally handles the 180-degree periodicity of oriented bounding boxes without discontinuities.

Performance is evaluated using Angle Accuracy, defined as the percentage of predictions where the model correctly identifies the exact 1-degree ground truth bin.

3.6 Results

3.6.1 Comparison of Pre-Training Approaches

Figure 3.4 presents qualitative results comparing predicted outputs with ground truth annotations for representative frames, demonstrating the model's ability to detect and localize regions of interest accurately.

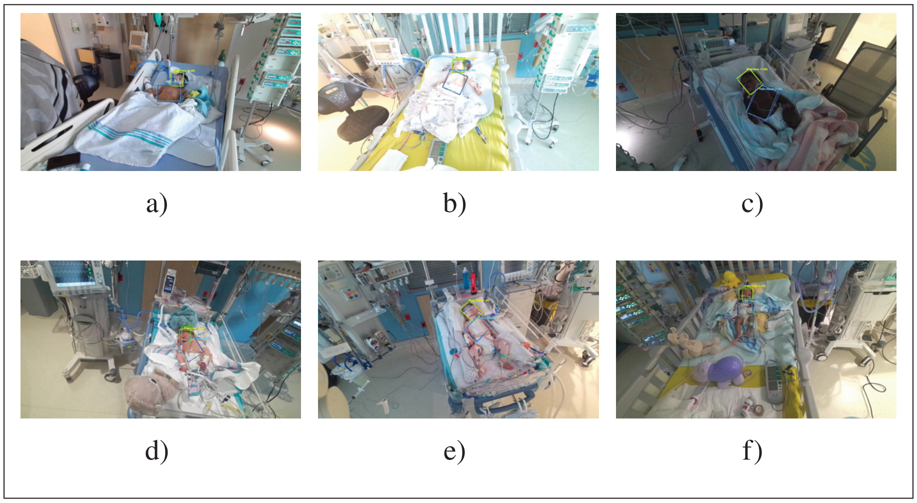


Figure 3.4 Subfigures (a)–(f) illustrate representative PICU frames from different patients under varying levels of occlusion and lighting conditions. Solid lines indicate ground truth annotations (green: face, blue: thorax), and dashed lines represent model predictions (cyan: face, yellow: thorax)

The comparative results of various pre-training strategies highlight the challenges of applying them in specialized domains such as the PICU, where data scarcity and distinct visual characteristics complicate transfer learning. Pediatric data differs significantly from adult datasets, and the presence of uncontrolled lighting, occlusions from medical equipment, varying poses, and intra-domain variability exacerbates the difficulty of learning robust representations.

As shown in Table 3.1, models trained from scratch or fine-tuned from supervised pre-training on Kinetics-400 struggle to converge effectively in this setting. Scratch training suffers from overfitting risks due to the simultaneous need to learn both low-level and domain-specific features, while Kinetics-400 pre-training, designed for action recognition, offers limited benefit for fine-grained spatial tasks like face and thoracoabdominal detection.

Table 3.1 Performance comparison of pre-training strategies using rotated IoU (rIoU), mean Average Precision (mAP) at different thresholds, and angle accuracy. The best results are bolded

Model	**rIoU**	**mAP@0.50**	**mAP@0.60**	**mAP@0.75**	**Angle**
From Scratch	0.85	0.56	0.41	0.22	0.25
K400	0.88	0.61	0.44	0.24	0.28
Teacher-Student	0.95	0.92	0.76	0.50	0.35
MAE (CHU-SJ Only)	0.94	0.94	0.82	0.65	0.37
PreTrain-MAE (Synth.)	**0.96**	**0.95**	**0.85**	**0.70**	**0.40**

This performance gap is reflected in the training and validation curves shown in Fig. 3.5. The scratch-trained model shows slow mAP improvement and high variance, while the Kinetics-400 pre-trained model converges faster but fails to generalize well due to domain mismatch. In contrast, masked self-supervised pre-training achieves smoother loss curves and consistently better mAP, indicating stronger domain alignment and robustness to PICU-specific challenges.

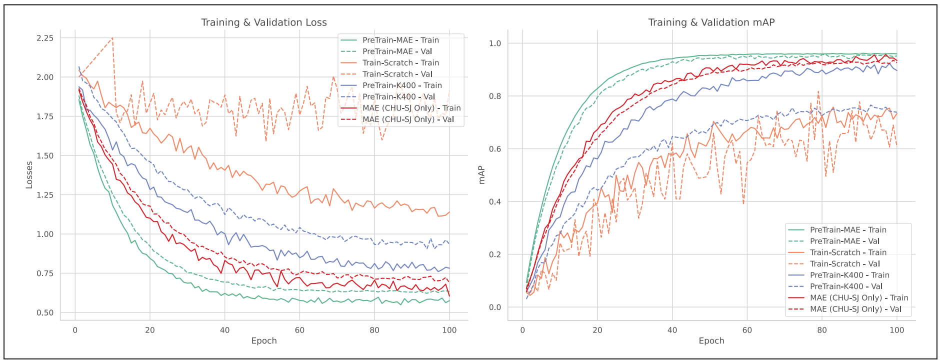


Figure 3.5 Comparison of training strategies. The PreTrain-MAE model, which was pre-trained using an augmented dataset of video clips generated from real clinical images, shows faster convergence and achieves a higher final mAP compared to the other baselines

While the Teacher–Student paradigm (UMT) outperforms both baseline models, its reliance on a CLIP-based teacher, which is trained on generic image distributions, limits its effectiveness in medical settings. In the PICU, where scenes often include tubes, blankets, and neonatal anatomy, this external guidance can introduce noise, misdirecting the student and destabilizing convergence. As shown in Table 2.3, UMT achieves only 0.50 $\text{mAP}_{@0.75}$, compared to 0.70 for the Masked Autoencoder (MAE). Unlike UMT, MAE operates without an external teacher and learns directly from the target data through high-ratio tube masking. This purely self-supervised strategy captures subtle appearance cues and motion patterns intrinsic to the PICU, enabling higher rIoU, mAP, and angle accuracy. Overall, MAE's domain-native learning approach proves more effective and reliable in clinical video environments characterized by data scarcity and complexity.

Furthermore, the effectiveness of our self-supervised approach is significantly enhanced by our data augmentation strategy. As illustrated in Figure 3.5, a direct comparison between the model pre-trained on clinical data alone (MAE (CHU-SJ Only)) and the one augmented with augmented clips (PreTrain-MAE) reveals a substantial performance gain. The model leveraging augmented sequences derived from real clinical data not only converges significantly faster but also achieves a higher final mAP and a lower training loss. This result, quantified in Table 3.1, provides direct evidence that our augmented data generation is a key contributor to the model's success, serving as an effective method to overcome data scarcity and improve generalization in this challenging clinical domain.

3.6.2 Model Validation

To verify the robustness of our best-performing model, we conducted 5-fold patient-wise cross-validation using the 373 training patients. The 112 test patients were held out exclusively for final evaluation and were not included in cross-validation. Patients were randomly divided into 5 folds of approximately 74-75 patients each, ensuring no patient appears in multiple folds. Each fold maintained similar distributions of age groups, occlusion severity, and recording conditions to avoid bias.

The cross-validation results in Table 3.2 demonstrate highly consistent performance across all folds, with minimal variance in key metrics (mAP@0.5: σ=0.006, rIoU: σ=0.005). This low variability indicates that our model generalizes well across different patient populations and is not overfitting to specific patient characteristics.

Table 3.2 Five-fold patient-wise cross-validation results for DST-Mamba model demonstrating stability across different patient subsets

Fold	**Train/Test**	**mAP@0.5**	**mAP@0.75**	**rIoU**	**Angle MAE**	**Temporal IoU**
1	298/75	0.948	0.682	0.952	0.41	0.94
2	299/74	0.962	0.705	0.961	0.39	0.96
3	298/75	0.951	0.693	0.958	0.42	0.95
4	300/73	0.957	0.701	0.963	0.40	0.95
5	297/76	0.960	0.698	0.955	0.38	0.94
Mean	-	0.956	0.696	0.958	0.40	0.95
±SD	-	±0.006	±0.009	±0.005	±0.015	±0.008
95% CI	-	(0.949, 0.963)	(0.685, 0.707)	(0.952, 0.964)	(0.382, 0.418)	(0.940, 0.960)

3.6.3 Comparison with State-of-the-Art Methods

To evaluate the effectiveness of our DST-Mamba approach, we compared it against established frame-based and video-based detection models, as summarized in Table 3.3. Frame-based models such as YOLOv8-m achieve high single-frame accuracy (0.892 mAP) but lack temporal coherence, resulting in unstable region-of-interest (ROI) tracking and jitter, which negatively impacts physiological signal estimation. This instability introduces motion artifacts into pixel-level signals, degrading the accuracy of downstream heart rate and respiratory estimation pipelines, particularly in neonates with subtle physiological cues. Furthermore, applying YOLOv8 across 16 frames incurs substantial computational load (634.6 GFLOPs) and high latency (243 ms). Incorporating DeepSORT (Bewley *et al.*, 2016) improves temporal consistency (0.82 temporal IoU) but further increases inference time. ViTDet(Li, Mao, Girshick & He, 2022) was also benchmarked in a per-frame configuration. While it achieved reasonable localization performance (0.60 mAP50-95), its high latency (115 ms) and parameter count (102M) reduce its suitability for real-time clinical deployment.

In contrast, video-based models such as I3D-FPN (Carreira & Zisserman, 2017) and TimeSformer (Bertasius *et al.*, 2021) either suffer from low detection accuracy (e.g., 0.360 mAP for I3D-FPN) or require significantly more parameters (up to 121M for TimeSformer) while still falling short in temporal stability. VideoMAE (ViT-B)(Tong *et al.*, 2022), achieves stronger performance (0.920 mAP, 0.90 temporal IoU) but with higher memory requirements and moderate latency. Notably, VideoMamba(Li *et al.*, 2024) achieves competitive performance across the board (0.940 mAP, 0.420 mAP50-95, 0.92 temporal IoU) while maintaining the lowest GFLOPs (5.71) and latency (35 ms), making it a strong benchmark for efficiency.

Table 3.3 Comprehensive comparison of models on accuracy, efficiency, and temporal stability for a 16-frame sequence

Model	**GFLOPs**	**Params (M)**	**Latency (ms)**	**mAP@0.5**	**mAP50-95**	**Temporal IoU**
Frame-based Models						
YOLOv8-m	634.6	26	243	0.892	0.445	0.75
YOLOv8-m + DeepSORT	634.6	28	352	0.926	0.465	0.82
ViTDet (per-frame)	270	102	115	0.69	0.60	0.80
Video-based Models						
I3D-FPN	174.38	35	180	0.360	0.200	0.50
TimeSformer	380	121	75	0.785	0.240	0.65
VideoMAE (ViT-B)	101.9	86	83	0.920	0.330	0.90
VideoMamba	5.71	54	35	0.940	0.420	0.92
DST-Mamba (Ours)	**7.56**	**73**	**43**	**0.960**	**0.620**	**0.95**

Our proposed DST-Mamba model directly addresses this challenge by achieving the highest detection accuracy (0.960 mAP) and the highest localization precision, with a 0.620 mAP50-95. Despite a modest increase in GFLOPs (7.56) and latency (43 ms) compared to VideoMamba, DST-Mamba offers a better overall trade-off between efficiency and accuracy. This advantage stems from its Divided Space–Time architecture, which enables specialized spatial and temporal learning without the overhead of joint attention. These findings suggest DST-Mamba is potentially suitable for real-time processing, pending clinical validation, balancing accuracy, stability, and efficiency critical for downstream physiological signal extraction.

3.6.4 Ablation Studies

3.6.4.1 Space-Time Mamba Architecture

We compared our sequential Mamba architecture with joint and parallel design variants, as shown in Table 3.4. While the joint model demonstrated slightly better computational efficiency (1.39 GFLOPs, 35 ms), it yielded lower detection accuracy (0.91 mAP). In contrast, the parallel architecture exhibited a severe performance drop (0.31 mAP), clearly indicating that independently modeling spatial and temporal features is insufficient for accurate region-of-interest (ROI) detection in PICU environments.

Table 3.4 Ablation study on Mamba-based architectures. Performance is evaluated for a 16-frame sequence at 224 × 224 resolution

Model Variant	**GFLOPs**	**Params (M)**	**Latency (ms)**	**mAP@0.5**
Parallel Space-Time	1.92	81	90	0.31
Joint Space-Time	1.39	54	35	0.91
Sequential (Ours)	**1.85**	**73**	**43**	**0.95**

Our proposed sequential design achieves the best balance, delivering the highest accuracy (0.95 mAP) with only a modest increase in computational cost (1.85 GFLOPs, 43 ms). These results strongly support our hierarchical design choice: extracting spatial features first, followed by temporal modeling, leads to more robust and reliable detection in complex clinical video settings. As illustrated in Figure 3.6, DST-Mamba converges faster and more smoothly than joint processing, ultimately reaching higher accuracy.

3.6.4.2 Comparative Analysis of Model Architectures

Table 3.5 compares the trade-off between computational efficiency and detection accuracy across three video models. While ViT achieves strong performance (0.91–0.95 mAP), it incurs significantly higher computational costs, ranging from 50.92 to 415.71 GFLOPs, and has a large parameter count (86.23M). In contrast, VideoMamba delivers competitive results (0.90–0.94 mAP) with substantially lower FLOPs (0.69–5.71) and a smaller model size (54.14M parameters).

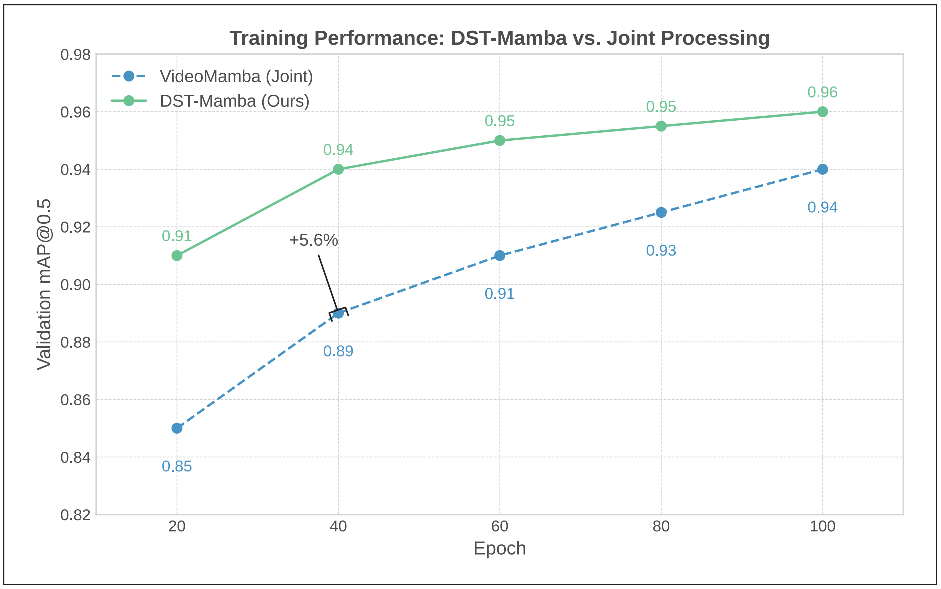


Figure 3.6 Training progression of DST-Mamba compared to the joint processing baseline. DST-Mamba achieves smoother convergence and consistently higher mAP

Notably, the Divided Space–Time Mamba model achieves the highest accuracy (0.96 mAP) at 16 frames and 640^2 resolution, while maintaining a moderate computational footprint (7.56 GFLOPs). This result highlights the effectiveness of its factorized design in achieving a favorable balance between accuracy and efficiency.

Table 3.5 Model performance and efficiency comparison. This table details the computational cost (FLOPs), model size (Parameters), and mean Average Precision (mAP) for different architectures, input resolutions, and frame counts

Model	**Frames**	**Input Size**	**FLOPs (G)**	**Parameters (M)**	**mAP**
ViT	8	224^2	50.92	86.23	0.91
ViT	16	224^2	101.85	86.23	0.93
ViT	16	640^2	415.71	86.23	0.95
VideoMamba	8	224^2	0.69	54.14	0.90
VideoMamba	16	224^2	1.39	54.14	0.91
VideoMamba	16	640^2	5.71	54.14	0.94
Divided Space–Time	8	224^2	0.93	73.65	0.89
Divided Space–Time	16	224^2	1.85	73.65	0.95
Divided Space–Time	16	640^2	7.56	73.65	**0.96**

3.6.4.3 Model Components

The results in Table I-1 demonstrate the impact of different architectural components on detection performance. Removing the angle loss function (Without Angle Loss) leads to the lowest performance across all metrics (IoU = 0.81, mAP@0.50 = 0.33, MAE = 0.37), confirming the importance of angle supervision for accurate localization and orientation. The fixed-angle variant (Without Orientation), which assumes vertical alignment, achieves high IoU (0.92) and zero angle error by design, but only moderate mAP (0.487), indicating limited adaptability to real-world orientation variability. Omitting the rotated IoU (Without rIoU) preserves high IoU (0.90) and mAP (0.92), but yields poor angular precision (MAE = 0.40), underscoring the importance of including orientation-aware overlap metrics. The depth-enabled model (With Depth) achieves the highest IoU (0.96) and mAP@0.50 (0.95), though with a slightly elevated MAE (0.52), suggesting that while depth improves spatial localization, it may increase complexity in estimating precise object orientation. These results underscore the necessity of integrating angle loss, orientation modeling, and rotated IoU, alongside depth information, for robust detection in complex clinical scenes.

3.6.4.4 Pre-training Masking Ratio

Table I-2 shows how different masking ratios affect both the efficiency of self-supervised pre-training and the performance on the downstream detection task. Among the configurations tested, a ratio of 80% provides the best trade-off: it achieves the highest mAP@0.5 (0.95), requires fewer fine-tuning epochs (70), and significantly reduces GPU memory usage (7.73 GB). This ratio introduces enough reconstruction difficulty to promote strong representation learning while preserving sufficient spatial context. In contrast, higher masking ratios (90–95%) reduce memory even further but slightly degrade accuracy, likely due to excessive information removal during pre-training. On the other end, lower ratios (50–70%) offer more visual cues but result in higher memory usage and slower convergence. These results confirm that 80% masking provides the best balance for our setting, optimizing resource usage while improving both learning speed and final detection accuracy. Figure 3.7 shows how different masking ratios affect pre-training and downstream detection. An 80% masking ratio provides the best trade-off, achieving the lowest pre-training loss and the highest fine-tuned mAP@0.5, while also reducing GPU memory usage and convergence time.

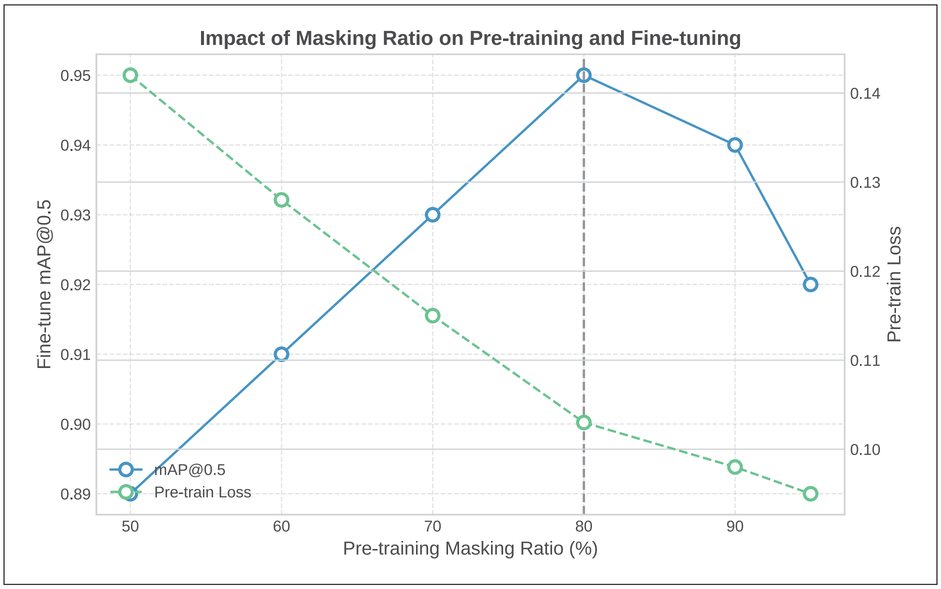


Figure 3.7 Impact of pre-training masking ratio on downstream detection performance. An 80% masking ratio achieves the lowest pre-training loss and highest mAP@0.5

3.6.4.5 Model Depth and Qualitative Results

Table I-3 summarizes the impact of encoder depth on model performance and computational efficiency. Increasing the number of layers from 4 to 12 leads to a substantial improvement in mAP@0.50 (from 0.88 to 0.95) and rIoU (from 0.90 to 0.96). However, further increasing the depth to 16 layers provides only marginal gains (0.96 mAP@0.50) while significantly increasing parameters and FLOPs. Thus, the 12-layer configuration (73.7 M parameters, 1.85 GFLOPs) offers the best trade-off between accuracy and efficiency.

3.6.5 Robustness to Clinical Occlusions

To better understand the limitations of our DST-Mamba model in clinical practice, we conducted a detailed error analysis across varying occlusion conditions. Test set predictions were manually categorized into four occlusion severity levels based on the percentage of anatomical region visibility, as shown in Table I-4. For the None category (less than 10% occlusion), the model achieved IoU of 0.96 and mAP@0.50 of 0.98, demonstrating robust performance under ideal visibility conditions. Light occlusion (10-25%), typically caused by medical tubes or monitoring leads, resulted in minimal performance degradation with IoU of 0.95 and mAP@0.50 of 0.96. Moderate occlusion (25-70%), including scenarios where blankets or oxygen masks partially covered the regions of interest, showed more substantial impact with IoU dropping to 0.88 and mAP@0.50 to 0.89. The mAP50-95 metric decreased from 0.61 to 0.52 between light and moderate occlusion levels, indicating reduced localization precision at higher IoU thresholds. Severe occlusion (exceeding 70% coverage) represented the primary failure mode for our model. Performance degraded significantly with IoU of 0.61 and mAP@0.50 of 0.58. The mAP50-95 dropped to 0.31, reflecting poor localization accuracy across all IoU thresholds. The 0.27 IoU drop from moderate to severe occlusion indicates substantial detection instability when most of the target region is obscured. These results demonstrate that while DST-Mamba maintains acceptable performance under partial occlusion, severe occlusion remains a significant challenge.

3.7 Discussion

Our results indicate that the Divided Space–Time (DST) Mamba architecture directly addresses the core obstacles of PICU video detection, setting a new state of the art for face and thoracoabdominal localization. By factorizing spatiotemporal modeling space-to-time, the model preserves temporal dynamics under motion and occlusions, while oriented boxes (OBBs) with circular smooth label (CSL) supervision explicitly handle orientation variability and camera skew, improving rIoU and reducing angle error. Domain-native self-supervised pretraining (MAE/UMT) mitigates data scarcity and domain shift, yielding higher mAP and more reliable convergence than training from scratch or Kinetics-400 finetuning. Integrating depth (RGB–D) further improves localization in occlusion-heavy, low-contrast scenes, with a measured compute trade-off and occasional angle-error increase that we report. Finally, the state-space formulation confers linear-time complexity and real-time, deployment-oriented efficiency (e.g., ~23 FPS at 16×640^2 with ~7.56 GFLOPs) while outperforming strong frame-wise and video baselines on accuracy and stability. Representative failure cases are presented in Figure I-1, illustrating reduced detection accuracy under severe occlusion, low illumination, and partial patient rotation.

3.7.1 From High-Accuracy Detection to Clinical Reliability

The extraction of physiological signals using non-contact methods like remote photoplethysmography (rPPG) is highly sensitive to the stability and consistency of the input ROI. Frame-based detectors process each frame independently, which introduces spatiotemporal jitter and increases inference time, limiting real-time applicability in clinical settings. In contrast, our DST-Mamba model achieves a high temporal IoU of 0.95, ensuring temporally coherent and stable ROIs. This stability reduces non-physiological motion noise and improves the signal-to-noise ratio (SNR), which is essential for reliable physiological monitoring. A mean Average Precision of 0.96 reflects both technical accuracy and clinically meaningful consistency in anatomical localization across frames. This level of performance minimizes gaps in region tracking, reducing the risk of signal disruption during rPPG or respiratory extraction. In critical care settings, even short detection lapses can lead to missed events or delayed alerts, making stable and accurate detection essential for continuous, high-fidelity monitoring. Additionally, the high rotational IoU ensures precise anatomical localization, preventing signal contamination from surrounding regions and reducing the risk of false alarms in intensive care environments.

3.7.2 Limitations & Future Work

The results of this study must be considered in the context of several key limitations. First, the model was developed and validated using data from a single institution. Its performance on data from other PICUs, which may differ in lighting conditions, camera configurations, and clinical protocols, remains untested. To our knowledge, no other publicly available PICU video datasets exist that capture both facial and thoracoabdominal regions simultaneously, which necessitated our reliance on single-center data. To address potential site-specific bias, our framework integrates multimodal RGB–Depth inputs, which are inherently less sensitive to lighting variations and camera-specific color calibration. In addition, we perform self-supervised pre-training on a heterogeneous corpus of over 50,000 video clips comprising both real PICU recordings and synthetically generated hospital scenarios with diverse illumination, contrast, and viewpoints. This domain-diverse pretraining strategy serves as an implicit form of cross-institutional regularization and enhances the model's robustness to unseen acquisition conditions. A multi-center validation is therefore a critical next step to assess whether the model can generalize. Although this study incorporates depth maps during training, real-time depth capture is not yet standard in most PICU monitoring systems. Future work should assess the feasibility and clinical value of integrating low-cost depth sensors at the bedside to enable robust 4D video analysis.

Second, the dataset is limited to 485 patients due to the practical and ethical challenges of data acquisition in pediatric critical care. The absence of public video datasets for this task necessitated our use of a synthetic data generation strategy. This strategy, while beneficial to pre-training performance in our experiments, is itself a limitation. We acknowledge that this method primarily provides rich spatial augmentation and does not capture true, physiologically relevant temporal dynamics. However, our results demonstrate that this spatial pre-training provides a crucial foundation, allowing the model to generalize much more effectively when subsequently fine-tuned on real clinical videos where it learns the relevant temporal patterns. Further work is required to determine if this method captures meaningful temporal dynamics or primarily provides spatial augmentation. Third, this study's scope is confined to the detection and tracking of anatomical regions. The work does not validate whether the improved detection metrics translate to more accurate downstream vital sign extraction. Establishing this link between technical performance and clinical utility is a crucial future step.

Finally, the model has not been tested in a live clinical workflow. Any claims regarding deployment readiness are premature, as real-world use requires prospective testing, integration with hospital IT systems, and navigating regulatory pathways.

Future work involves integrating our DST-Mamba architecture with vital sign extraction algorithms for prospective clinical validation against contact monitors. To address data scarcity and foster collaboration, we will release our open-source code and augmented data generation methodology upon publication. The code will be made publicly available at: https://github.com/mkbensalah/Divided-Space-Time-Mamba.

3.8 Conclusions

In this paper, we presented the Divided Space–Time Video Mamba framework for medical video detection in Pediatric Intensive Care Unit (PICU) environments. By decoupling spatial and temporal processing, our approach achieves high accuracy (0.95 mAP@0.50) while maintaining computational efficiency. The incorporation of masked autoencoder pre-training further improves performance, reaching 0.96 rIoU and 0.95 mAP@0.50, and shows improved performance on our single-center dataset. Additionally, integrating depth information enhances the model's robustness to occlusions and variable lighting conditions. The efficiency of DST-Mamba supports its integration into real-time clinical systems. With 7.56 GFLOPs and 73M parameters, the model processes 16-frame inputs at 640 × 640 resolution in 43 ms, achieving 23 FPS. Its linear complexity ensures predictable scalability with respect to input resolution and sequence length, unlike transformer-based models that scale quadratically. This makes it suitable for deployment on edge devices such as bedside monitors or portable diagnostic tools in the PICU. Furthermore, the divided space–time structure promotes interpretability by isolating spatial and temporal contributions. This separation facilitates the analysis of detection failures and helps verify consistency across frames. The use of oriented bounding boxes produces rotation-aware outputs that align with clinical requirements, especially in respiratory monitoring where thoracic orientation directly influences signal quality. In addition, saliency-based visualizations or attribution maps can be employed to reveal the specific regions within each frame that most influence the model's predictions. Such visual feedback allows clinicians to verify that the model attends to relevant anatomical features, rather than being misdirected by medical equipment, patient coverings, or shadows. Future work will focus on extending DST-Mamba toward cross-device generalization by evaluating its robustness across different RGB-D sensors and acquisition settings, and by incorporating domain-adaptation techniques to mitigate sensor-specific variability. We will also investigate lightweight pruning, quantization, and token-reduction strategies to enable efficient deployment on embedded and bedside monitoring systems. This study was conducted under institutional ethical approval from CHU Sainte-Justine, with all video data processed within secure research servers. Future work will explore privacy-preserving edge AI and federated learning frameworks to ensure patient data remain local while enabling continuous, on-device model adaptation for real-time bedside use.

CHAPTER 4

NON-CONTACT PHYSIOLOGICAL MONITORING IN PEDIATRIC INTENSIVE CARE UNITS VIA ADAPTIVE MASKING AND SELF-SUPERVISED LEARNING

Mohamed Khalil Ben Salah[1] , Philippe Jouvet[2] , Rita Noumeir[1]

[1] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[2] Research Center at CHU Sainte-Justine Hospital, University of Montreal,
3175 Chem. de la Côte-Sainte-Catherine, Montréal, Québec, Canada H3T 1C5

Article submitted to « IEEE Journal of Biomedical and Health Informatics (JBHI) » in November 2025.

4.1 Abstract

Continuous monitoring of vital signs in Pediatric Intensive Care Units (PICUs) is essential for early detection of clinical deterioration and effective clinical decision-making. However, contact-based sensors such as pulse oximeters may cause skin irritation, increase infection risk, and lead to patient discomfort. Remote photoplethysmography (rPPG) offers a contactless alternative to monitor heart rate using facial video, but remains underutilized in PICUs due to motion artifacts, occlusions, variable lighting, and domain shifts between laboratory and clinical data. We introduce a self-supervised pretraining framework for rPPG estimation in the PICU setting, based on a progressive curriculum strategy. The approach leverages the VisionMamba architecture and integrates an adaptive masking mechanism, where a lightweight Mamba-based controller assigns spatiotemporal importance scores to guide probabilistic patch sampling. This strategy dynamically increases reconstruction difficulty while preserving physiological relevance. To address the lack of labeled clinical data, we adopt a teacher–student distillation setup. A supervised expert model, trained on public datasets, provides latent physiological guidance to the student. The curriculum progresses through three stages: clean public videos, synthetic occlusion scenarios, and unlabeled videos from 500 pediatric patients. Our framework achieves a 42% reduction in mean absolute error relative to standard masked autoencoders and outperforms PhysFormer by 31%, reaching a final MAE of 3.2 bpm. Without explicit region-of-interest

extraction, the model consistently attends to pulse-rich areas and demonstrates robustness under clinical occlusions and noise.

4.2 Introduction

Contactless monitoring of vital signs in Pediatric Intensive Care Units (PICUs) is essential to ensure patient safety and support early clinical intervention. Traditional contact-based sensors, such as electrocardiograms (ECGs) and pulse oximeters, are widely used but remain suboptimal for fragile neonates. These sensors can irritate sensitive skin and increase the risk of infection (Khanam, Perera, Al-Naji, Gibson & Chahl, 2021). Remote photoplethysmography (rPPG), which estimates blood volume pulse (BVP) from subtle changes in facial color, provides a non-contact alternative. Since the COVID-19 pandemic, rPPG has attracted growing interest across telemedicine, fitness applications, and clinical monitoring. In the PICU setting, its unobtrusive nature offers the potential to reduce patient discomfort, minimize caregiver handling, and enable continuous monitoring. A reliable rPPG solution that provides heart rate monitoring, could improve pediatric critical care by eliminating the need for adhesive or invasive sensors.

Despite its potential, deploying rPPG models in clinical settings remains challenging. The PICU environment presents multiple sources of signal degradation, including occlusions (e.g., tubes, caregivers), variable lighting, motion artifacts, and high inter-subject variability. Moreover, supervised learning approaches are limited by the scarcity of labeled clinical data, as video acquisition in PICUs is ethically sensitive and logistically constrained. Public datasets such as UBFC-rPPG and PURE are collected under controlled conditions and do not generalize well to real-world clinical environments. This domain mismatch often results in poor performance when models trained on lab data are applied in clinical practice (Villarroel *et al.*, 2019).

To address these challenges, we introduce a self-supervised pretraining framework for rPPG estimation, specifically designed for the PICU context. Our framework is trained on an Institutional Review Board (IRB)-approved dataset of 500 pediatric patient videos and follows a curriculum learning strategy. The training progresses in three stages: initial learning on clean public datasets, followed by synthetic occlusion augmentation, and culminating in large-scale pretraining on real PICU data. This staged approach improves model robustness to clinically relevant artifacts and environmental variability.

Our framework employs a masked autoencoder backbone inspired by VideoMAE (Tong *et al.*, 2022), and introduces two core innovations. First, we incorporate an Adaptive Masking Network (AMN) that replaces conventional random masking. The AMN uses a Mamba-based controller trained via policy gradient reinforcement learning to assign importance scores to each spatiotemporal patch. By sampling binary masks that occlude highly informative regions, typically those containing strong pulsatile signals, the model is forced to develop more generalizable and physiologically meaningful representations. This mechanism allows the model to implicitly attend to key facial regions, without requiring explicit region-of-interest (ROI) extraction, and enhances generalization under challenging conditions such as occlusion and noise.

Second, we integrate a multi-objective training scheme to guide learning. The main generative objective combines Mean Squared Error (MSE) with a pixel-level Pearson correlation term to ensure accurate temporal modeling of BVP dynamics. While this promotes global spatiotemporal learning, it does not guarantee alignment with physiological targets. To introduce this inductive bias, we adopt a teacher–student distillation approach. A supervised expert model provides high-quality reference signals, and the student network is trained to align its predictions with these targets. This process reinforces the extraction of clinically meaningful features, even in cases of partial visibility or degraded video quality.

Our key contributions are as follows:

- We introduce a contactless vital-sign monitoring framework for PICUs that lowers skin lesions, and enables continuous physiological monitoring.
- We propose a self-supervised rPPG framework that integrates masked autoencoding with Pearson correlation and physiological distillation to learn robust pulsatile dynamics from facial videos.

- We present an adaptive masking strategy using a Mamba-based controller, which dynamically selects spatiotemporal patches to improve robustness against occlusions, motion artifacts, and variable lighting.
- We validate our method on a large IRB-approved dataset of 500 pediatric patients and demonstrate its effectiveness for rPPG and heart rate estimation under realistic PICU conditions.

4.3 Related Works

4.3.1 Remote Photoplethysmography (rPPG)

Conventional rPPG methods typically follow a two-stage pipeline involving facial region detection and signal extraction based on RGB intensity variations that reflect blood volume pulse (BVP) dynamics. Early signal processing techniques such as Independent Component Analysis (ICA) (Poh *et al.*, 2010a), CHROM (De Haan & Jeanne, 2013), and Plane-Orthogonal-to-Skin (POS) (Wang *et al.*, 2017a) were developed to isolate pulsatile components while mitigating motion and illumination artifacts. Although effective under controlled conditions, these handcrafted approaches rely on fixed assumptions about lighting and skin reflectance, which limit their applicability in clinical environments such as the PICU, where motion, occlusion, and sensor noise are common. These constraints have motivated the shift toward deep learning models that offer improved robustness through data-driven learning.

Initial deep learning approaches for rPPG estimation employed 2D convolutional neural networks (CNNs). Some introduced motion-based attention mechanisms (Chen & McDuff, 2018), while others incorporated video magnification (Qiu *et al.*, 2019a) or emphasized spatial features linked to pulse signals (Špetlík *et al.*, 2018). Subsequent models, including EfficientPhys (Liu *et al.*, 2022a) and Multi-Task Temporal Shift Convolutional Attention Network (MTTS-CAN) (Liu *et al.*, 2020), proposed architectural refinements to better capture temporal dynamics. However, 2D CNNs remain limited in modeling spatiotemporal dependencies. This led to the use of 3D CNNs, such as rPPGNet (Yu *et al.*, 2019c) and PhysNet (Yu *et al.*, 2019a), which process entire video volumes. Further improvements were introduced through attention-based architectures (Hu *et al.*, 2021a,b) designed to enhance signal localization.

Other strategies include hybrid CNN-RNN models (Niu *et al.*, 2020a; Hu *et al.*, 2019; Huang *et al.*, 2021a) to model temporal dependencies, and meta-learning methods (Lee *et al.*, 2020a) to improve generalization across subjects and domains. Despite these advances, 3D CNNs often remain more effective in clinical applications due to their ability to jointly model spatial and temporal features.

More recently, transformer-based models have been adopted for rPPG estimation (Vaswani *et al.*, 2017). Vision Transformers (ViT) (Dosovitskiy *et al.*, 2020) inspired specialized designs such as TransPPG (Kang *et al.*, 2024) and PhysFormer (Yu *et al.*, 2022), which incorporate motion awareness and periodicity constraints. TimeSformer (Bertasius *et al.*, 2021) introduced divided space-time attention, and rPPGTR (Zhang *et al.*, 2023a) extended this to combine local CNN features with global attention for robust signal reconstruction. However, these approaches still depend on supervised training and large-scale labeled datasets, which are often unavailable in clinical contexts.

Our work builds on these developments by proposing a self-supervised learning framework that leverages large-scale unlabeled PICU data. The model is trained using a staged curriculum designed to capture the variability of real-world clinical conditions without requiring manual annotations.

4.3.2 Self-Supervised Learning for Video Understanding

Self-supervised learning (SSL) has emerged as a promising direction for learning video representations from unlabeled data, particularly in medical applications where annotation is both costly and privacy-sensitive (Xu *et al.*, 2019). Pretext tasks such as contrastive learning (Chen *et al.*, 2020b; Grill *et al.*, 2020; Caron *et al.*, 2021) and masked reconstruction (He *et al.*, 2022) have demonstrated strong performance in visual domains.

Among these, generative approaches based on Masked Image Modeling (MIM) have shown particular effectiveness. The Masked Autoencoder (MAE) (He *et al.*, 2022) employs an asymmetric encoder–decoder structure to reconstruct masked patches, enabling efficient learning of semantically meaningful features. This approach has been extended to video with VideoMAE (Tong *et al.*, 2022), which applies high masking ratios (90–95%) across spatiotemporal tokens using a tube-based masking pattern. By leveraging temporal redundancy, VideoMAE achieves robust pretraining while maintaining computational efficiency through a lightweight Vision Transformer (ViT) encoder. This design highlights the importance of pretraining directly on domain-specific video datasets to improve downstream task performance.

Recent extensions such as Unmasked Teacher (UMT) (Li *et al.*, 2023b) build on this paradigm through a teacher–student framework that guides the masking process. Instead of randomly selecting patches, UMT uses teacher-derived attention maps to mask less informative regions, thereby improving semantic alignment and enhancing representation learning efficiency.

Despite these advances, most existing SSL methods rely on static or random masking strategies that do not adapt to the content variability in clinical videos. In tasks such as rPPG estimation, spatial heterogeneity caused by motion, occlusion, and physiological signal variability requires a more targeted masking approach. To address this, we propose an adaptive masking mechanism driven by a lightweight Mamba-based policy network. Unlike fixed masking schemes, our method learns to identify and suppress visually physiologically regions, guiding the reconstruction process toward clinically informative areas. This improves feature learning under the complex and variable conditions characteristic of the PICU environment.

4.3.3 SSL for Remote Photoplethysmography

SSL has emerged as an effective strategy to address the scarcity and high acquisition cost of physiological annotations in remote photoplethysmography. Recent SSL frameworks have explored diverse objectives to improve generalization under conditions involving motion, occlusion, illumination changes, and noise (Hasan, Faridee, Ahmed & Roy, 2022; Zhang, Sun, Ma & Jia, 2024; Wang, Ahn & Kim, 2022b; Xiao *et al.*, 2024).

Several approaches have focused on contrastive learning, where the objective is to distinguish between temporally or spatially aligned positive pairs and unaligned negatives. Gideon et al. (Gideon & Stent, 2021) introduced a fully self-supervised method that relies on weak priors in the frequency and temporal domains, removing the need for labels. Contrast-Phys (Sun & Li, 2022) and its extension Contrast-Phys+ (Sun & Li, 2024) applied 3D convolutional backbones with contrastive objectives to extract physiological representations from facial videos. These models achieved improved robustness to motion and noise, performing on par with supervised baselines.

In parallel, generative SSL techniques based on masked autoencoding have demonstrated advantages in learning structural and periodic signal information. rPPG-MAE (Liu *et al.*, 2024b) used a masked reconstruction objective to model the self-similarity of BVP signals, showing improved resilience to motion and illumination variability. Notably, the authors emphasized that pretraining performance depends more on dataset quality than size. Yue et al. (Yue, Shi & Ding, 2023) proposed a frequency-aware SSL method that combines spatial and frequency domain augmentations with a learnable frequency transform module, enabling better encoding of signal periodicity in the absence of labels.

SSL methods incorporating teacher–student structures or pseudo-labeling have also gained traction. PhySU-Net employed pseudo-labels derived from conventional signal extraction pipelines to guide a masked image reconstruction task. Similarly, Li et al. (Li & Yin, 2023) proposed a co-rectification strategy to improve training stability in the presence of noisy pseudo-supervision. These works aim to combine handcrafted signal priors with learned representations.

Transformer-based SSL models have recently gained momentum for capturing long-range spatiotemporal dependencies. RS-rPPG (Savic & Zhao, 2024b) leveraged temporal augmentation and a transformer backbone, achieving improved performance over earlier contrastive models.

ST-Phys (Cao *et al.*, 2024) demonstrated strong robustness to occlusion and noise using a lightweight architecture. SimPPG (Bhattachrjee, Li, Xia & Xu, 2023) introduced a region-based non-contrastive framework using positive pairs from the same subject, allowing effective physiological representation learning without explicit labels. TransPhys (Wang, Sun, Hao, Pan & Jia, 2023) further advanced this line of work by applying contrastive transformer pretraining for global feature learning, outperforming existing SSL benchmarks in rPPG estimation.

These studies highlight the growing potential of SSL for rPPG under uncontrolled conditions. However, most prior work relies on uniform masking or augmentation schemes that ignore the spatial and temporal variability present in clinical data. Such models typically treat all input regions as equally informative, lacking mechanisms to focus on physiologically relevant areas. In addition, current frameworks often omit structured curricula or physiological guidance, limiting their capacity to generalize to the complex visual dynamics of PICU recordings.

To address these challenges, we introduce a unified self-supervised framework that integrates multiple components previously studied in isolation. Specifically, we propose an adaptive masking strategy driven by a lightweight Mamba-based controller, a teacher–student distillation mechanism that transfers physiological priors from a supervised model, and a progressive curriculum learning schedule that bridges the gap between controlled datasets and real-world PICU data. This combination improves model robustness to occlusions, motion artifacts, lighting variability, and limited annotation, which are characteristic of intensive care settings.

4.3.4 State Space Models (SSMs)

Transformers (Vaswani *et al.*, 2017) have become the dominant architecture for sequential data modeling in natural language processing and computer vision due to their ability to capture global dependencies through multi-head self-attention. However, their quadratic complexity with respect to sequence length presents scalability challenges, particularly for high-resolution images and long video sequences.

State Space Models (SSMs) have emerged as a scalable alternative for long-sequence modeling, offering linear computational complexity. Structured State Space sequence models (S4) (Gu *et al.*, 2021) demonstrated that reparameterized state space representations can capture long-range dependencies efficiently. S4 introduced low-rank updates and diagonal stabilization by normalizing parameter matrices. Subsequent variants, including S5 (Smith *et al.*, 2022), H3 (Fu *et al.*, 2022), and GSS (Mehta *et al.*, 2022), have explored trade-offs in stability, expressivity, and computational efficiency.

Mamba (Gu & Dao, 2023) recently introduced a data-dependent SSM mechanism that enables selective processing of input sequences. By making transition matrices input-dependent and optimizing for parallel execution, Mamba achieves efficient long-range modeling with linear complexity. It has demonstrated strong performance across large-scale datasets and has been applied to diverse tasks (Pióro *et al.*, 2024).

In the visual domain, Mamba has been adapted for spatial and spatiotemporal modeling. Vision Mamba (ViM) (Zhu *et al.*, 2024) extends Mamba to 2D using bidirectional scans for improved spatial representation. VMamba (Liu *et al.*, 2024c) introduces a four-directional Selective Scan (SS2D) for enhanced contextual propagation. EfficientVMamba (Pei *et al.*, 2025) reduces computational cost via atrous sampling, while LocalMamba (Huang *et al.*, 2024) focuses on local context through windowed scanning and spatial-channel attention. Mamba-based backbones have also been adopted in medical image analysis, including segmentation and detection tasks (Ruan *et al.*, 2024; Liu *et al.*, 2024a).

These developments show the potential of SSM-based models to process long visual sequences with improved scalability. In this work, we leverage Vision Mamba to model clinical video sequences, where capturing temporally extended patterns and ensuring computational efficiency are critical.

4.4 Methodology

4.4.1 Overview

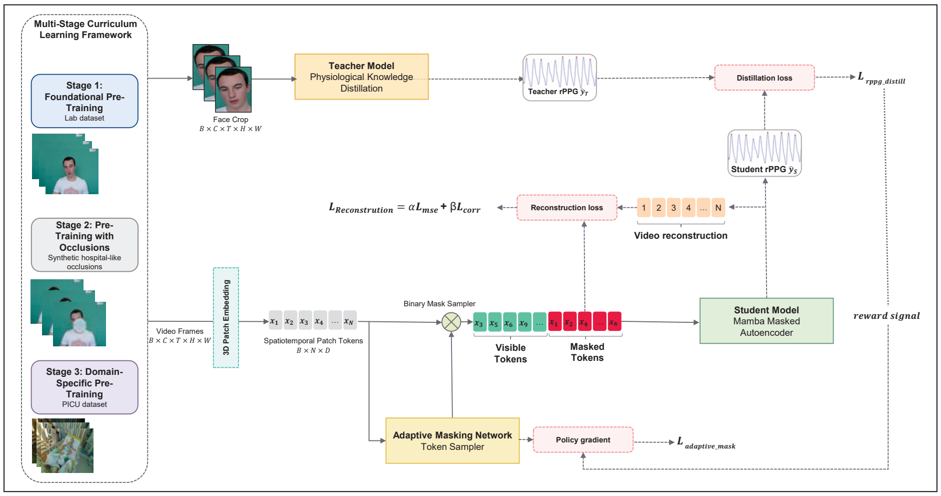

Multi-Stage Curriculum Learning Framework
Stage 1: Foundational Pre-Training
Lab dataset
Stage 2: Pre-Training with Occlusions
Synthetic hospital-like occlusions
Stage 3: Domain-Specific Pre-Training
PICU dataset
Face Crop
$B \times C \times T \times H \times W$
Teacher Model
Physiological Knowledge Distillation
Teacher rPPG $\hat{y}_T$
Distillation loss
$L_{rppg\_distill}$
Student rPPG $\hat{y}_S$
$L_{Reconstrution} = \alpha L_{mse} + \beta L_{corr}$
Reconstruction loss
1 2 3 4 ... N
Video reconstruction
Video Frames
$B \times C \times T \times H \times W$
3D Patch Embedding
x_1 x_2 x_3 x_4 ... x_N
Spatiotemporal Patch Tokens
$B \times N \times D$
Binary Mask Sampler
x_3 x_5 x_6 x_9 ... x_1 x_2 x_4 ... x_N
Visible Tokens
Masked Tokens
Student Model
Mamba Masked Autoencoder
reward signal
Adaptive Masking Network
Token Sampler
Policy gradient
$L_{adaptive\_mask}$

Figure 4.1 Overview of our curriculum-based framework for robust rPPG estimation. The student model learns by reconstructing masked patches and predicting the rPPG signal from visible tokens. Training is guided by an expert PhysMamba teacher model and a learnable Adaptive Masking Network (AMN) optimized via policy gradient reinforcement

This work proposes a self-supervised framework for robust remote photoplethysmography estimation in Pediatric Intensive Care Unit environments. The method combines adaptive masked autoencoding with physiological knowledge distillation to improve signal robustness under common challenges such as occlusion, lighting variation, and patient motion.

As shown in Figure 4.1, the architecture includes three components: a student model, a teacher model, and a learnable Adaptive Masking Network (AMN).

The student model is optimized on two tasks. It receives spatiotemporal tokens from an input video, a subset of which is masked by the AMN. The model must: (i) reconstruct the masked patches via a lightweight decoder in a self-supervised setting, and (ii) estimate the rPPG signal from the unmasked tokens through a regression head. The PhysMamba teacher provides the reference rPPG signal, enabling physiological knowledge transfer.

The AMN is the core novelty of this framework. Rather than using random masking, it learns to select and mask informative spatiotemporal tokens based on their importance. It consists of lightweight Mamba blocks that assign importance scores to tokens and apply differentiable Gumbel-Top-K sampling (Jang, Gu & Poole, 2016) to choose which tokens to mask. AMN is trained via a policy gradient strategy using the student's rPPG loss as a reward. The masking policy is rewarded when it increases the student's prediction error, pushing the student to extract more robust and generalizable features.

To improve generalization, we adopt a three-stage curriculum. The model is first trained on clean laboratory videos, then on synthetically occluded samples, and finally on real PICU data from 500 pediatric patients.

4.4.2 VisionMamba Student Model

As illustrated in Figure 4.2, the student architecture consists of three main components: a patch tokenizer, an encoder, and a decoder head.

4.4.2.1 Patch Tokenizer

The student model processes an input video $\mathbf{X} \in \mathbb{R}^{C \times T \times H \times W}$, where C is the number of channels, T the number of frames, and $H \times W$ the spatial resolution. A 3D convolutional patch embedding layer with kernel size and stride (t, h, w) partitions the input into non-overlapping tubelets, generating patch tokens $\mathbf{X}_p \in \mathbb{R}^{N \times D}$, where N is the number of tokens and D the embedding dimension. To retain spatial and temporal ordering, fixed sinusoidal positional embeddings are added to the patch tokens.

4.4.2.2 Encoder Block

The Adaptive Masking Network generates a binary mask that retains a subset of visible tokens $\mathbf{X}_{\text{vis}} \in \mathbb{R}^{K \times D}$, where $K < N$. These visible tokens are passed through an encoder composed of L stacked bidirectional Mamba blocks with residual connections, followed by a normalization layer. The output is linearly projected to match the decoder's input dimension and combined with learnable mask tokens corresponding to the masked positions. This combined sequence is used for reconstruction in the decoder.

4.4.2.3 Decoder Head

The decoder receives the visible token representations, projected into the decoder embedding dimension, along with the learnable mask tokens corresponding to the masked positions. Fixed positional embeddings are added to both sets, which are then concatenated to form the full decoder input. This sequence is processed by a stack of D Mamba blocks, followed by a normalization layer. The decoder head, implemented as a linear projection, reconstructs only the masked tokens, yielding patch-level outputs for self-supervised learning.

In parallel, the model includes an rPPG prediction head. It aggregates both visible and masked token embeddings using mean pooling and feeds the result into a multilayer perceptron (MLP) with dropout and non-linear activation to predict the rPPG waveform of length T. This dual objective encourages the model to learn both spatial reconstruction and physiologically relevant temporal dynamics.

4.4.3 Adaptive Masking Network

Standard masked autoencoders typically apply static or random masking schemes that lack content awareness, often removing physiologically informative regions or retaining uninformative ones. To address this limitation, we introduce an Adaptive Masking Network, a learnable policy module that selectively masks the most informative spatiotemporal tokens in an adversarial manner. By exposing the student only to low-signal or ambiguous tokens, the AMN increases task difficulty and encourages the extraction of robust features relevant to physiological dynamics.

The AMN operates on the patch embeddings $\mathbf{x} \in \mathbb{R}^{B\times N\times D}$, derived from a 3D convolutional projection of the input video $\mathbf{x}_{\text{video}} \in \mathbb{R}^{B\times C\times T\times H\times W}$. Instead of relying on attention-based architectures, the AMN leverages a lightweight stack of Mamba blocks to efficiently capture long-range spatial and temporal dependencies. The output is passed through a shallow MLP to compute per-token importance scores.

To guide masking toward relevant regions, a spatial prior bias is added, favoring central facial areas. Gumbel noise is introduced to ensure sampling variability, and a Top-K operation selects the top $(1 - r)\%$ tokens, where r is the target masking ratio. The remaining tokens are masked. The final visibility mask is binary and differentiable, as illustrated in Figure 4.2.

This mask is applied to the token sequence, allowing the student to process only the retained tokens, while the masked positions are replaced with a learned embedding. The AMN is optimized not by conventional backpropagation but via policy gradient reinforcement learning. The reward signal is defined by the student's rPPG prediction loss, with higher errors corresponding to more challenging masks. This setup drives the AMN to occlude high-signal regions, such as the forehead or cheeks, forcing the student to learn from sparse and degraded observations. The full procedure for generating the mask is described in Algorithm 4.1.

Formally, given the importance logits $\mathbf{l} \in \mathbb{R}^{B\times N}$ and the sampled binary mask $\mathbf{m} \in \{0, 1\}^{B\times N}$, the log-probability of the selected masking action is computed from the softmax distribution over $\mathbf{l}$. The policy gradient loss is defined as:

$$\mathcal{L}_{\text{PG}} = -\mathbb{E}\left[\left(\mathcal{L}_{\text{rPPG}} - \bar{\mathcal{L}}\right) \cdot \log p(\mathbf{m})\right] \tag{4.1}$$

Algorithm 4.1 Adaptive Masking Network (AMN)

```
Algorithm: Adaptive Masking Network (AMN)
Input: Input video $\mathbf{X} \in \mathbb{R}^{B\times C\times T\times H\times W}$, AMN parameters $\theta$, temperature $\tau$, mask ratio $\rho$
Output: Binary mask $\mathbf{M} \in \{0,1\}^{B\times N}$, importance scores $\mathbf{S} \in \mathbb{R}^{B\times N}$
Initialization:;
$N \leftarrow (T/t_s) \times (H/p_s) \times (W/p_s)$;                  /* Number of patches */
$N_{vis} \leftarrow \lfloor N \times (1-\rho) \rfloor$;               /* Number of visible patches */
Initialize spatial bias $\mathbf{B}_{spatial}$ centered on face region;
Stage 1: Patch Embedding;
$\mathbf{P} \leftarrow \text{PatchEmbed}(\mathbf{X})$;                /* $\mathbf{P} \in \mathbb{R}^{B\times N\times D}$ */
$\mathbf{P} \leftarrow \mathbf{P} + \text{PositionalEncoding}(N, D)$;
Stage 2: Importance Score Computation;
$\mathbf{H} \leftarrow \mathbf{P}$;
for each MambaBlock in AMN do
    $\mathbf{H} \leftarrow \text{MambaBlock}(\mathbf{H})$;            /* Process through Mamba SSM */
end for
$\mathbf{H} \leftarrow \text{LayerNorm}(\mathbf{H})$;
$\mathbf{S}_{raw} \leftarrow \text{ImportanceHead}(\mathbf{H})$;      /* $\mathbf{S}_{raw} \in \mathbb{R}^{B\times N\times 1}$ */
$\mathbf{S} \leftarrow \text{Squeeze}(\mathbf{S}_{raw}) + \mathbf{B}_{spatial}$;
Stage 3: Differentiable Mask Sampling;
$\mathbf{S}_{logits} \leftarrow \mathbf{S}/\tau$;                    /* Temperature scaling */
$\mathbf{S}_{logits} \leftarrow \text{Clamp}(\mathbf{S}_{logits}, -10, 10)$;
$\mathbf{G} \leftarrow -\log(-\log(\mathbf{U} + \epsilon) + \epsilon)$;   /* Gumbel noise, $\mathbf{U} \sim \text{Uniform}(0,1)$ */
$\mathbf{S}_{perturbed} \leftarrow \mathbf{S}_{logits} + \mathbf{G}$;
$idx_{vis} \leftarrow \text{TopK}(\mathbf{S}_{perturbed}, N_{vis})$;
$\mathbf{M} \leftarrow \mathbf{1}_{B\times N}$;                       /* Initialize mask (1 = masked) */
$\mathbf{M}[idx_{vis}] \leftarrow 0$;                                 /* Set visible patches to 0 */
return $\mathbf{M}, \mathbf{S}_{logits}$
```

where $\mathcal{L}_{\text{rPPG}}$ is the signal-level distillation loss between the student and teacher rPPG waveforms, $\bar{\mathcal{L}}$ is a batch-wise baseline for variance reduction, and $p(\mathbf{m})$ is the probability of the sampled visible tokens.

The AMN is jointly trained with the student model but does not receive direct gradients from the decoder or the regression head. Instead, it learns indirectly through its effect on the student's performance, effectively acting as a hard instance generator that strengthens feature robustness.

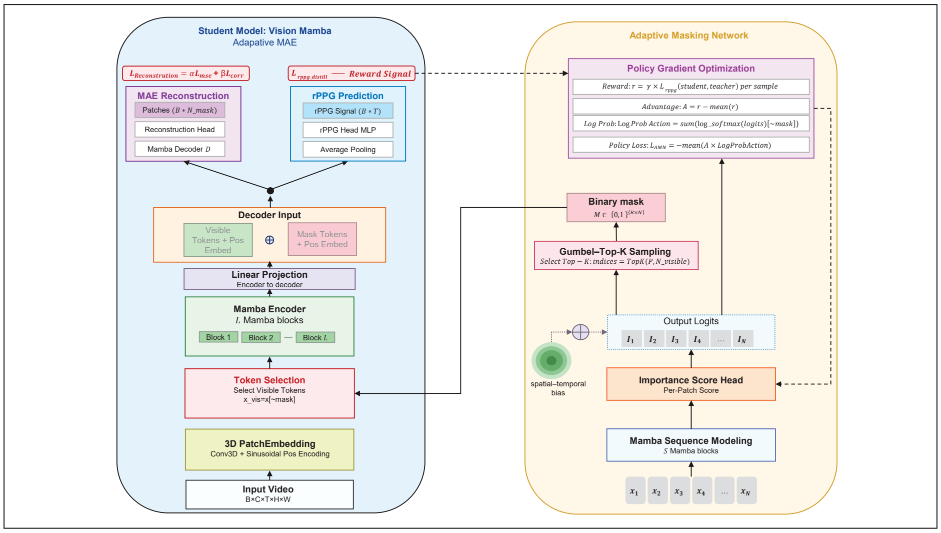

Student Model: Vision Mamba
Adapative MAE
$L_{Reconstrution} = \alpha L_{mse} + \beta L_{corr}$
$L_{rppg\_distill}$ — Reward Signal
MAE Reconstruction
Patches $(B * N\_mask)$
Reconstruction Head
Mamba Decoder D
rPPG Prediction
rPPG Signal $(B * T)$
rPPG Head MLP
Average Pooling
Decoder Input
Visible Tokens + Pos Embed
⊕
Mask Tokens + Pos Embed
Linear Projection
Encoder to decoder
Mamba Encoder
L Mamba blocks
Block 1
Block 2
Block L
Token Selection
Select Visible Tokens
x_vis=x[~mask]
3D PatchEmbedding
Conv3D + Sinusoidal Pos Encoding
Input Video
B×C×T×H×W
Adaptive Masking Network
Policy Gradient Optimization
$Reward: r = \gamma \times L_{rppg}(student, teacher)$ per sample
$Advantage: A = r - mean(r)$
$Log\ Prob: \mathrm{Log}\ Prob\ Action = sum(\log\_softmax(logits)[\sim mask])$
$Policy\ Loss: L_{AMN} = -mean(A \times LogProbAction)$
Binary mask
$M \in \{0,1\}^{\{B\times N\}}$
Gumbel–Top-K Sampling
$Select\ Top-K: indices = TopK(P, N\_visible)$
Output Logits
I_1 I_2 I_3 I_4 … I_N
spatial–temporal bias
Importance Score Head
Per-Patch Score
Mamba Sequence Modeling
S Mamba blocks
x_1 x_2 x_3 x_4 … x_N

Figure 4.2 Detailed architecture of the Adaptive Masking Network (AMN). The AMN computes token importance scores using Mamba blocks and selects a visible subset via Gumbel–Top-K sampling. Policy gradient optimization uses the rPPG distillation loss as reward to update the AMN. The student reconstructs the masked tokens and predicts the rPPG waveform

4.4.4 Physiological Knowledge Distillation

Although self-supervised masked reconstruction supports general representation learning from unlabeled clinical videos, it lacks explicit physiological guidance, which is critical for capturing the fine-grained temporal dynamics required for accurate rPPG estimation. To overcome this limitation, we integrate a distillation mechanism in which a pre-trained expert model, PhysMamba (Luo *et al.*, 2024), serves as a fixed teacher during pretraining.

PhysMamba is a supervised rPPG model trained on clean, face-cropped clean datasets. It processes DiffNormalized frames, where each frame is temporally normalized relative to its neighbors to enhance pulsatile information and suppress motion and lighting variability. Within our framework, PhysMamba provides physiological supervision by generating high-fidelity reference signals, which guide the student during masked autoencoding. While knowledge distillation is commonly used in general vision tasks, its use with a domain-specific Mamba model for self-supervised rPPG pretraining in clinical video remains largely unexplored. This design embeds physiological priors into the learned representations and improves temporal consistency in the presence of clinical artifacts. The student incorporates a dedicated rPPG prediction head, which regresses the waveform from pooled token embeddings. Specifically, global average pooling is applied across both visible and masked spatiotemporal tokens before decoding. This provides the rPPG head with access to the overall temporal context required for signal reconstruction.

To enforce alignment between the student's predictions and the teacher-generated reference signals, we apply a signal-level distillation loss based on negative Pearson correlation:

$$\mathcal{L}_{\text{distill}} = 1 - \rho(\hat{\mathbf{y}}_{\text{student}}, \mathbf{y}_{\text{teacher}}) \tag{4.2}$$

where $\rho(\cdot, \cdot)$ denotes the Pearson correlation coefficient between the predicted and teacher-generated rPPG waveforms. No feature-level alignment is enforced, avoiding over-constraining the latent space and ensuring that learning remains focused on waveform morphology and temporal coherence.

This loss serves a dual purpose: it provides a direct physiological supervision signal and simultaneously functions as a reward for the Adaptive Masking Network. Coupling the distillation loss with the masking policy gradient encourages the AMN to suppress informative patches, thereby increasing task difficulty and reinforcing robust temporal learning in the student.

Through this distillation mechanism, the student acquires the ability to infer physiologically accurate waveforms even under severe occlusion or partial visibility, mirroring realistic challenges encountered in PICU settings.

4.4.5 Training Objectives and Optimization

The training framework jointly optimizes three complementary objectives that promote robust learning of physiological features. The total loss combines masked reconstruction, physiological signal distillation, and masking policy optimization through a weighted formulation. To decouple feature learning from masking behavior, we adopt a dual-optimizer setup.

4.4.5.1 Masked Reconstruction Objective

The reconstruction loss focuses exclusively on masked tokens and consists of two components designed to preserve both spatial fidelity and temporal coherence. Let $\mathbf{X} \in \mathbb{R}^{B\times C\times T\times H\times W}$ denote the input video and $\mathbf{M} \in \{0,1\}^{B\times N}$ the binary mask generated by AMN, where N is the number of spatiotemporal patches. The overall reconstruction loss is:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{corr}}. \tag{4.3}$$

4.4.5.1.1 Pixel-Level MSE

This term penalizes local reconstruction errors by minimizing per-pixel differences over masked patches:

$$\mathcal{L}_{\text{pixel}} = \frac{1}{|\mathcal{M}|} \sum_{i\in\mathcal{M}} \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2, \tag{4.4}$$

where $\mathcal{M}$ denotes the masked indices, and $\hat{\mathbf{x}}_i$ and $\mathbf{x}_i$ are the reconstructed and ground-truth patches, respectively.

Minimizing pixel-wise error alone is insufficient for rPPG estimation. It often leads to temporally static outputs that ignore the subtle brightness oscillations encoding physiological signals.

4.4.5.1.2 Global Pearson Correlation

To address this limitation, we introduce a correlation-based term that captures global temporal dynamics across masked regions. The correlation is computed over flattened patch sequences:

$$\mathcal{L}_{\text{corr}} = 1 - \frac{1}{B}\sum_{i=1}^{B}\rho_i, \tag{4.5}$$

where ρ_i is the Pearson correlation between predicted and ground-truth sequences for sample i.

This loss integrates several physiological priors essential for waveform reconstruction. First, temporal coherence is enforced by ensuring masked patches span the full duration T, covering multiple cardiac cycles. This allows the model to preserve phase-consistent oscillations caused by pulsatile blood flow. Second, spatial consistency is encouraged, reflecting the fact that facial perfusion induces coordinated brightness changes across multiple regions such as the cheeks, nasal bridge, and forehead. Finally, the correlation term promotes preservation of dynamic range. Pearson correlation is insensitive to amplitude but penalizes shape mismatch, discouraging flat or low-variance reconstructions even when MSE is minimized. These constraints guide the model toward physiologically meaningful predictions.

4.4.5.2 Physiological Distillation Objective

To embed physiological priors during pretraining, we apply knowledge distillation from a fixed teacher model trained on clean, face-cropped videos. The teacher outputs reference rPPG signals from DiffNormalized input. The student predicts an rPPG waveform from globally pooled token embeddings. The distillation loss is computed as:

$$\mathcal{L}_{\text{distill}} = 1 - \rho(\hat{\mathbf{y}}_s, \mathbf{y}_t), \tag{4.6}$$

where $\hat{\mathbf{y}}_s$ and $\mathbf{y}_t$ are the normalized student and teacher signals, respectively. This term aligns temporal dynamics and encourages physiologically plausible predictions without restricting intermediate features.

4.4.5.3 Policy Gradient for Adaptive Masking

The Adaptive Masking Network is trained to suppress signal-rich patches, making the reconstruction and regression tasks more challenging. Given the per-token importance scores $\mathbf{l} \in \mathbb{R}^{B \times N}$, we apply differentiable Top-K sampling by perturbing these scores with Gumbel noise (Jang *et al.*, 2016):

$$\tilde{\mathbf{l}}_i = \frac{\mathbf{l}_i}{\tau} + \mathbf{g}_i, \quad \mathbf{g}_i \sim \text{Gumbel}(0, 1), \tag{4.7}$$

where τ denotes the temperature. The top $(1 - r) \times N$ tokens with the highest scores are selected as visible patches. This approach follows the Gumbel–Softmax trick, which enables a continuous relaxation of the Top-K operation and allows gradient flow through the sampling step. As a result, the network can learn content-aware masking strategies while maintaining stochasticity during training.

The AMN is optimized using a policy gradient formulation, where the student's rPPG distillation loss serves as a task-specific reward. The policy loss is defined as:

$$\mathcal{L}_{\text{PG}} = -\mathbb{E}_{\mathcal{V}} \left[A(\mathcal{V}) \cdot \sum_{j \in \mathcal{V}} \log p_j \right], \tag{4.8}$$

where $p_j = \text{softmax}(\mathbf{l})_j$ is the selection probability of token j, and the advantage function is:

$$A(\mathcal{V}) = w_{\text{rppg}} \cdot (\mathcal{L}_{\text{distill}} - b), \tag{4.9}$$

with $w_{\text{rppg}} = 2.0$ and b a baseline computed per batch to reduce variance. This training setup encourages the AMN to discover masking patterns that degrade the student's signal prediction performance, thereby forcing the model to learn more generalizable and temporally consistent representations. A complete summary of the AMN update procedure is provided in Algorithm 4.2.

Algorithm 4.2 Policy Gradient Training for Adaptive Masking

```
Algorithm: Policy Gradient Training for Adaptive Masking
Input: Student model $f_s$, Teacher model $f_t$, AMN $g_\theta$, rPPG weight $\beta$, loss weight $\alpha$
Output: Updated AMN parameters $\theta$
Forward Pass:;
$\mathbf{M}, \mathbf{S} \leftarrow g_\theta(\mathbf{X})$;                                  /* Generate adaptive mask */
$\mathbf{X}_{masked} \leftarrow \text{ApplyMask}(\mathbf{X}, \mathbf{M})$;
$\hat{\mathbf{y}}_{student} \leftarrow f_s(\mathbf{X}_{masked})$;                         /* Student rPPG prediction */
$\mathbf{y}_{teacher} \leftarrow f_t(\mathbf{X}_{face})$;                                 /* Teacher rPPG */
Reward Computation:;
$\mathcal{L}_{distill} \leftarrow 1 - \rho(\hat{\mathbf{y}}_{student}, \mathbf{y}_{teacher})$;     /* $\rho$ is Pearson correlation */
$R \leftarrow \mathcal{L}_{distill}$;                                                   /* Reward signal */
$b \leftarrow \text{Mean}(R)$;                                                          /* Baseline */
$A \leftarrow (R - b) \times \beta$;                                                    /* Advantage with rPPG weight */
Policy Gradient Loss:;
$\pi(\mathbf{M}|\mathbf{S}) \leftarrow \text{LogSoftmax}(\mathbf{S})$;                  /* Policy distribution */
$\log p(\mathbf{M}) \leftarrow \sum_{i \in \neg \mathbf{M}} \pi_i$;                     /* Log probability of visible patches */
$\mathcal{L}_{adaptive} \leftarrow -\text{Mean}(\log p(\mathbf{M}) \times \text{detach}(A)) \times \alpha$;
Parameter Update:;
$\theta \leftarrow \theta - \eta_a \nabla_\theta \mathcal{L}_{adaptive}$;
return $\theta$
```

4.4.5.4 Optimization Strategy

To ensure stable convergence and disentangle representation learning from policy learning, we employ two independent AdamW optimizers. The parameters of the student model, denoted by θ_s, are updated according to:

$$\theta_s^{(t+1)} = \theta_s^{(t)} - \eta_s \nabla_{\theta_s} (\lambda_{\text{mae}} \mathcal{L}_{\text{recon}} + \lambda_{\text{dist}} \mathcal{L}_{\text{distill}}), \tag{4.10}$$

where the learning rate η_s is set to 10^{-4}, the weight decay coefficient is $\lambda_w = 0.05$, and gradients are clipped such that $\|\mathbf{g}\|_2 \leq 1.0$ to prevent instability during backpropagation.

The parameters of AMN, denoted by ϕ, are optimized separately using the policy gradient loss:

$$\phi^{(t+1)} = \phi^{(t)} - \eta_{\text{amn}} \nabla_{\phi} \mathcal{L}_{\text{PG}}, \tag{4.11}$$

with a learning rate $\eta_{\text{amn}} = 10^{-5}$. Gradients from $\mathcal{L}_{\text{PG}}$ do not propagate through the student network. This design maintains a clear separation between the learning of visual representations and the adaptation of masking behavior.

4.5 Experimental Protocol

4.5.1 Datasets

4.5.1.1 Public Datasets for Pretraining

We adopt a curriculum-based pretraining strategy using three publicly available rPPG datasets before transitioning to the clinical PICU domain. These datasets introduce increasing variability in lighting, motion, and acquisition setup, enabling the model to progressively learn robustness across controlled and dynamic scenarios.

- **UBFC-rPPG** (Bobbia *et al.*, 2019): consists of 42 videos from 42 subjects recorded under indoor controlled conditions using a Logitech C920 HD Pro webcam. Each video is approximately 90 seconds long and recorded at 30 fps with a resolution of 640 × 480 pixels. Participants remained relatively still while engaging in a time-constrained mental arithmetic task to induce natural heart rate variation. Ground truth PPG signals were acquired using a CMS50E pulse oximeter sampled at 60 Hz.
- **VIPL-HR** (Niu, Han, Shan & Chen, 2018): includes 2,378 RGB videos from 107 subjects captured under varying illumination, spontaneous head movements (e.g., talking, gaze shifts),

and with multiple acquisition devices (Logitech C310, Intel RealSense F200, Huawei P9), resulting in heterogeneous resolutions and frame rates (25–30 fps). Some sessions were conducted post-exercise to diversify heart rate ranges. Ground truth BVP signals were collected via a CONTEC CMS60C sensor. The dataset introduces variability that facilitates generalization to out-of-distribution domains.

- **ECG-Fitness** (Špetlík *et al.*, 2018): contains 204 videos from 17 subjects (14 male, 3 female; aged 20 to 53 years) performing various physical activities such as talking, rowing, and exercising on ellipticals or bikes. Videos were recorded at 30 fps and 1280 × 720 resolution under three lighting setups: natural light, 400 W halogen, and 30 W LED. ECG signals were synchronously recorded at 256 Hz. The presence of substantial motion artifacts makes this dataset critical for pretraining in dynamic conditions relevant to clinical scenarios.

4.5.1.2 Clinical Dataset: The CHU-SJ PICU Collection

For domain adaptation and final evaluation, we use a large-scale clinical dataset acquired at the Pediatric Intensive Care Unit of CHU Sainte-Justine (CHU-SJ) (Boivin *et al.*, 2023). Data collection was approved by the institutional ethics committee (protocols #2019-2035 and #2016-1242), with informed parental consent obtained prior to recording. The dataset comprises 30-second video segments from 500 pediatric patients across a broad age range (1 month to 18 years), representing diverse ethnic backgrounds and skin tones.

Recordings were performed using a multimodal acquisition system centered around the Microsoft Azure Kinect DK, capturing high-resolution RGB video at 3840 × 2160 resolution and 30 fps. Synchronized physiological signals including PPG, ECG, and oxygen saturation (SpO_2) were simultaneously collected from Philips IntelliVue MX800 bedside monitors at 125 Hz.

This dataset presents considerable challenges due to the complexity of real-world clinical settings. Videos frequently include occlusions, such as caregiver interventions and medical devices including oxygen masks (6.0%), ventilator tubes (3.2%), blankets (5.8%), and hats (4.4%). The cohort includes both spontaneously breathing and mechanically ventilated patients, with varying levels of supplemental oxygen. These characteristics provide a rigorous benchmark for evaluating the robustness and deployment feasibility of rPPG models in hospital environments.

4.5.2 Training Details

All experiments were implemented in PyTorch and executed on a server with four NVIDIA A100 GPUs.

4.5.2.1 Data Preprocessing and Input Configuration

For the teacher model, facial regions were detected using YOLO5Face. Bounding boxes were enlarged by a factor of 1.2 to encompass the full face. Each frame was resized to 224×224 and normalized using ImageNet statistics. Every training sample includes $T = 128$ consecutive frames, equivalent to approximately 4.3 seconds of video at 30 fps. During training, heart rate was estimated via FFT peak detection within the 0.7–3.0 Hz physiological band. For evaluation on 30-second clips, a sliding window of 128 frames was used, with HR computed per window. Window-level HR values were smoothed using a median filter (kernel size 5) and averaged across windows to obtain one estimate per video. Ground-truth PPG was downsampled to 30 Hz and temporally aligned with the video using synchronized timestamps.

The student model uses a tokenizer with a tubelet size of $(t, h, w) = (2, 16, 16)$, producing a sequence of $N = 12{,}544$ spatiotemporal tokens per clip.

4.5.2.2 Hospital Occlusion Simulation

Stage 2 incorporates a Hospital Occlusion Simulator replicating seven common occlusion types observed in the PICU. These include: (1) medical tubes as Bézier curves (thickness 3–8 pixels), (2) oxygen masks as ellipses covering 70% of the face width, (3) medical tape as rotated rectangles (10–25 × 15–40 pixels), (4) hands as irregular ellipses entering from frame edges, (5) shadows via gradient overlays (intensity 0.3 to 0.7), (6) equipment obstructing 15–30% of the frame width, and (7) blanket edges modeled using cubic splines.

Each occlusion persists for 20–80% of the clip with 70% temporal consistency. Following the curriculum strategy, occlusion probability increases linearly from 0% to 50% between epochs 50 and 150, and spatial coverage increases from 10% to 40% of the face.

Colors match typical clinical appearances: grayscale (100–180) for equipment, high-intensity white (200–255) for tubes, and skin-tone or blue for gloves and hands. This augmentation increases robustness to visual disruptions encountered in clinical practice, supporting domain adaptation in Stage 3.

4.5.2.3 Model and Optimization Hyperparameters

The VisionMamba student encoder consists of 12 bidirectional Mamba blocks (dim 768). The decoder includes 8 Mamba blocks (dim 384), and the Adaptive Masking Network (AMN) uses 4 Mamba layers.

Self-supervised pretraining spanned 2,400 epochs across three curriculum stages. Stage 1 (Foundational Pretraining) used public datasets for 600 epochs to learn general representations. Stage 2 (Robustness Pretraining) introduced synthetic clinical occlusions and ran for 1,000 epochs. Stage 3 (Domain-Specific Pretraining) involved 800 epochs on PICU data to align with clinical distribution.

Architectural and training hyperparameters were selected via systematic grid search and convergence analysis on a validation subset of the PICU datasets. We optimized the masking ratio, architecture depth, and stage durations to ensure stable learning and computational efficiency. Early stopping criteria were also considered to prevent overfitting during extensive training phases.

Masking ratios from 50% to 95% (in 5% steps) were tested. A 75% ratio was optimal: lower values (50–70%) led to trivial reconstructions (MSE < 0.01), while higher values (80–95%) caused instability in 30% of training runs. A 75% ratio also ensured visibility of at least one facial region in 90% of frames and matched average occlusion severity in PICU videos (see Figure 4.9).

For architecture depth, using more than 12 encoder layers yielded marginal gains (MAE improvement < 0.2 bpm) at a 25% memory cost increase. Thus, we selected 12 encoder, 8 decoder, and 4 AMN layers. Increasing AMN depth by more than 4 layers led to 40% longer training without significant masking accuracy improvement (only a 3% gain in importance variance).

A dual-optimizer strategy was used. The student was trained using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with an initial learning rate of 1×10^{-4} and cosine annealing, following a 40-epoch warm-up. AMN used a separate AdamW optimizer with a constant learning rate of 1×10^{-5}. Weight decay of 0.05 was applied to all trainable parameters except biases and normalization terms. Gradient clipping was enforced with a maximum L_2 norm of 1.0 for both optimizers to stabilize adversarial learning.

4.5.2.4 Supervised Fine-tuning

Following self-supervised pre-training, we perform supervised fine-tuning on annotated PICU data to enable accurate absolute heart rate estimation. From the full cohort of 500 patients, we selected a labeled subset of 200 patients. This subset was divided into 160 patients for training and 40 for testing. To enhance generalization, we conducted 5-fold patient-wise cross-validation across the 200 annotated patients. Each fold used 160 patients for training and 40 for testing, ensuring strict patient-level separation. All reported results reflect the average across folds, with 95% confidence intervals.

To avoid data leakage, we enforced rigorous patient-level data partitioning. The 40 test patients were entirely excluded from all phases of training, including self-supervised pretraining. Specifically, during Stage 3 pretraining, we limited the PICU subset to 460 patients by excluding the 40 test patients and the 200 labeled patients reserved for evaluation. We also applied patient-independent clip splitting, ensuring that all 30-second clips from a single patient were assigned exclusively to either the training or test set. This protocol prevents the model from overfitting to patient-specific patterns and ensures generalizability across unseen subjects.

During supervised fine-tuning, the AMN is deactivated and bypassed. All spatial tokens are forwarded to the encoder without masking. Only the student encoder and the rPPG prediction head are updated during this stage. The loss function jointly aligns predicted and reference PPG signals and penalizes heart rate error:

$$\mathcal{L}_{\text{supervised}} = 1 - \rho(\hat{\mathbf{y}}, \mathbf{y}_{\text{GT}}) + \lambda_{\text{HR}} \cdot |\text{HR}(\hat{\mathbf{y}}) - \text{HR}(\mathbf{y}_{\text{GT}})| \tag{4.12}$$

where $\rho(\cdot, \cdot)$ denotes the Pearson correlation between the predicted signal $\hat{\mathbf{y}}$ and ground truth $\mathbf{y}$GT, and HR$(\cdot)$ computes heart rate using FFT peak detection in the 0.7–3.0 Hz frequency range. The loss weighting factor λ_{HR} is set to 0.5.

Fine-tuning was performed using AdamW with a learning rate of 5×10^{-5}, set ten times lower than that used during pretraining. A cosine annealing schedule was applied across 100 epochs, with early stopping based on validation MAE. This phase refines the learned representations for precise clinical estimation while maintaining the robustness acquired through self-supervised learning.

4.5.3 Evaluation metrics

We report three quantitative metrics to assess heart rate estimation performance: mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient (R). All metrics are computed on a per-clip basis and averaged across the test set.

1. Mean absolute error MAE: It is calculated by taking the average absolute difference between the predicted values $HR_{predict}$ and the ground truth values $HR_{reference}$:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |HR^{i}_{predict} - HR^{i}_{reference}|. \tag{4.13}$$

2. Root mean squared error RMSE: It measures the average magnitude difference between the predicted value of heart rate and the actual one

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(HR^{i}_{predict} - HR^{i}_{reference})^{2}}. \tag{4.14}$$

3. The Pearson correlation coefficient R : It measures the linear correlation between the two signals $rPPG_{predict}$ and $PPG_{reference}$:

$$R = \frac{\mathrm{cov}(rPPG_{predict}, PPG_{reference})}{\mathrm{std}(rPPG_{predict}) \cdot \mathrm{std}(PPG_{reference})}. \tag{4.15}$$

4.6 Results and Discussion

4.6.1 Comparison With State-of-the-Art Methods on the PICU Dataset

Table 4.1 Performance Comparison on PICU Test Set

Method	MAE (bpm)	RMSE (bpm)	MPE (%)	R
CHROM(De Haan & Jeanne, 2013)	15.8	19.2	18.3	0.42
POS(Wang *et al.*, 2017a)	14.2	17.8	16.7	0.48
PhysNet(Yu *et al.*, 2019b)	8.7	11.3	10.2	0.71
EfficientPhys(Liu *et al.*, 2022a)	7.9	10.4	9.3	0.74
PhysMamba(Luo *et al.*, 2024)	7.2	9.8	8.5	0.77
MTTS-CAN(Liu *et al.*, 2020)	6.5	8.9	7.6	0.81
PhysFormer(Yu *et al.*, 2022)	5.8	8.2	6.8	0.84
Contrast-Phys(Sun & Li, 2022)	11.3	14.2	13.1	0.58
rPPG-MAE(Liu *et al.*, 2024b)	9.6	12.4	11.2	0.66
Ours (w/o fine-tuning)	10.3	17.6	14.2	0.67
Ours (w/ fine-tuning)	**3.2**	**5.4**	**3.8**	**0.91**

As shown in Table 4.1, the fine-tuned model achieves a MAE of 3.2 bpm and RMSE of 5.4 bpm on the PICU test set, surpassing both traditional and deep learning approaches. Compared to PhysFormer (MAE: 5.8 bpm), the proposed method reduces the error by 44.8% and achieves a 50.8% improvement over MTTS-CAN (MAE: 6.5 bpm). The Pearson correlation increases from 0.84 to 0.91, indicating closer alignment between the predicted and reference waveforms.

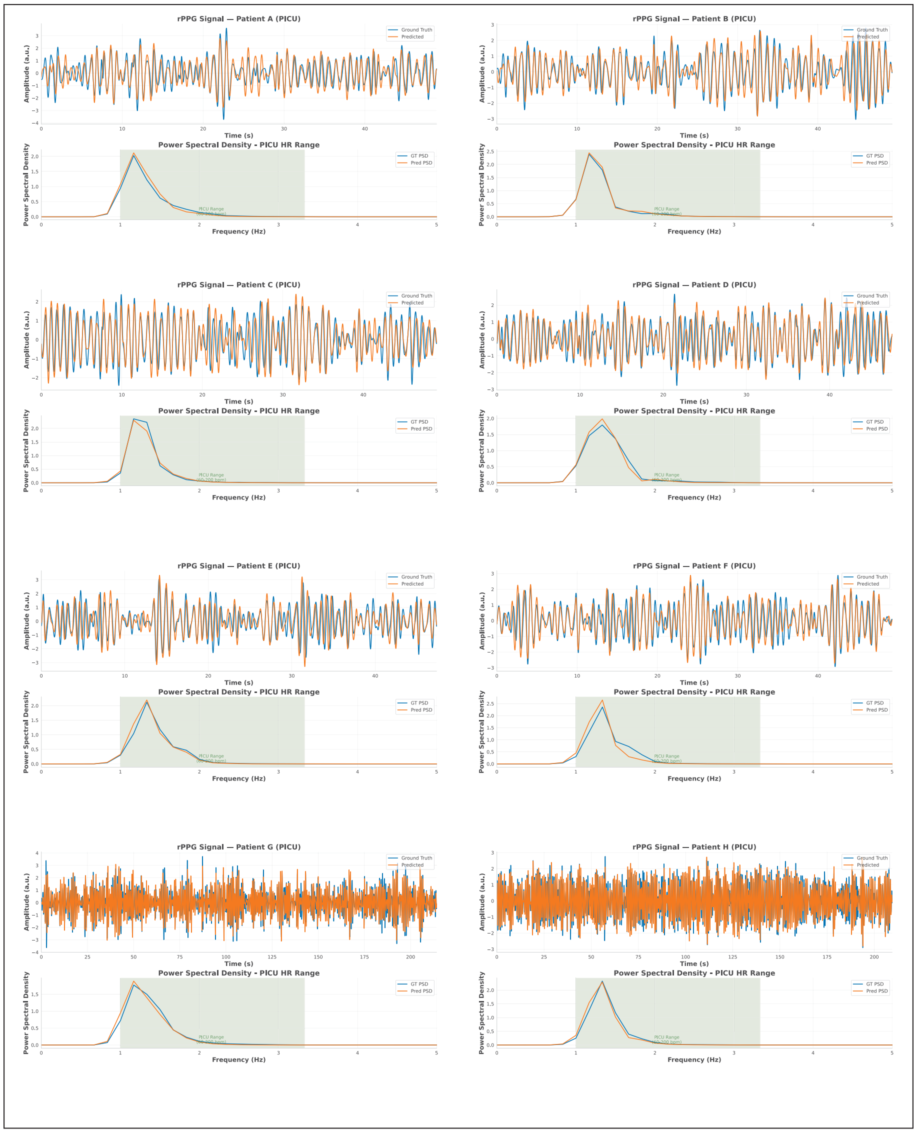


Figure 4.3 Qualitative signal reconstruction for six PICU patients. Each subplot shows the reconstructed rPPG waveform and corresponding ground truth, visualized in both the time domain and the frequency domain through Power Spectral Density (PSD) curves

Without supervised fine-tuning, the self-supervised model already attains a MAE of 10.3 bpm, outperforming Contrast-Phys (11.3 bpm) and approaching the performance of early supervised architectures. Fine-tuning on 160 labeled PICU patients further reduces the error from 10.3 bpm to 3.2 bpm, representing a relative improvement of 68.9%. This demonstrates that the pretraining stage provides a strong initialization that captures physiologically meaningful temporal features, while supervised adaptation refines these representations for clinical precision.

This combination yields robust performance in the presence of domain-specific challenges such as motion, occlusion, and poor illumination. Fine-tuning on 160 annotated patients provides a substantial reduction in error and confirms that the pretraining strategy learns meaningful spatiotemporal dynamics from large-scale unlabeled PICU data.

4.6.2 Qualitative Analysis of rPPG Signal Reconstruction

Figure 4.3 shows qualitative results from eight patients in the PICU cohort, comparing predicted rPPG signals and corresponding power spectral density (PSD) curves to ground truth references. In the time domain, the model preserves the pulsatile morphology and beat-to-beat variability. In the frequency domain, the PSD curves exhibit sharp, localized peaks aligned with the ground truth heart rate, indicating robustness to motion, occlusion, and subject variability.

These observations align with the quantitative results in Table 4.1, and further support the model's ability to recover physiologically meaningful signals without relying on fixed regions of interest or landmark-based supervision. The framework maintains performance even under occlusion, leveraging spatiotemporal dynamics learned during training.

The final two examples in Figure 4.3 correspond to recordings longer than 210 seconds. The reconstructed signals preserve regular pulsatile structure across the full duration, and the associated PSD curves remain confined within the expected pediatric heart rate range (60–200 bpm, or 1.0–3.3 Hz). This illustrates the model's capacity to support continuous monitoring where long-term signal stability is critical.

Figure 4.4 highlights cases where the model compensates for reference signal artifacts. In many PICU recordings, the contact-based oximeter shows flat segments or discontinuities due to disconnection or motion. Despite these interruptions, the predicted rPPG signals remain regular, and physiologically plausible.

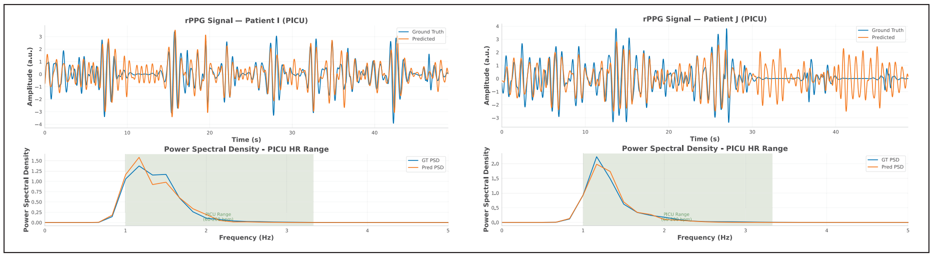


Figure 4.4 rPPG signal reconstruction on PICU patients showing resilience to ground truth sensor artifacts

For example, in Patient I (35–40 s) and J (20–25 s), the reference PPG signal exhibits abrupt degradation, while the model output maintains continuous pulse waveforms. This highlights a clinical advantage: unlike contact sensors that are prone to physical failure, the proposed method provides stable monitoring even under challenging acquisition conditions.

In the frequency domain, the reconstructed PSD curves retain sharp peaks within the expected heart rate range, confirming that the model preserves rhythmic content even when the reference signal is unreliable or missing.

4.6.3 Performance Across Different Patient Demographics

Table 4.2 Performance Stratified by Patient subgroups

Patient Group	**N**	**MAE (bpm)**	**RMSE (bpm)**	**R**
Neonates (0-1 month)	35	3.8	5.9	0.82
Infants (1-12 months)	77	3.4	5.5	0.87
Toddlers (1-3 years)	27	3.1	5.2	0.91
Children (3-12 years)	12	2.9	5.0	0.93
Adolescents (12-18 years)	8	1.7	3.8	0.95
Type I-II (Light)	14	2.8	4.9	0.92
Type III-IV (Medium)	18	3.2	5.4	0.91
Type V-VI (Dark)	8	3.9	6.1	0.88
Mechanically Ventilated	24	3.5	5.7	0.87
Non-Ventilated	16	2.7	4.9	0.93

As summarized in Table 4.2, the model achieves consistent performance across all age subgroups. The MAE decreases progressively with age, from 3.8 bpm in neonates to 1.7 bpm in adolescents. This trend indicates stable generalization across physiological variability and developmental motion patterns.

When stratified by skin tone, the model shows slightly higher error for Fitzpatrick Types V–VI (3.9 bpm) compared to Types I–II (2.8 bpm), a difference of 1.1 bpm. This aligns with prior findings on reduced signal-to-noise ratio in rPPG estimation due to melanin absorption (Dasari, Prakash, Jeni & Tucker, 2021; Talukdar, de Deus & Sehgal, 2023; Nowara, McDuff & Veeraraghavan, 2020).

Performance also differs by ventilation status. Mechanically ventilated patients present an MAE of 3.5 bpm, while non-ventilated patients achieve 2.7 bpm. The presence of medical devices such as tubing or adhesive patches may explain this difference by introducing occlusions or artifacts. Despite these subgroup differences, Pearson correlation remains high (0.87 to 0.95), confirming robust temporal alignment with ground truth heart rate.

4.6.4 Robustness to Clinical Occlusions

We assessed model performance under common occlusion scenarios encountered in the PICU environment. As shown in Figure 4.5, the most frequent sources of obstruction include medical devices and bedding. Oxygen masks and cloth coverings are particularly prominent, occluding an average of 6.1% and 5.9% of the image area, respectively. These occlusions often affect key facial regions required for accurate rPPG estimation.

Despite these challenges, the proposed model maintains stable performance across all occlusion categories and consistently outperforms baseline methods (Table I-4). In cases of severe occlusion (greater than 70% of the facial area), our method achieves a mean absolute error (MAE) of 7.2 bpm, compared to 13.3 bpm with PhysFormer and 16.8 bpm with MTTS-CAN. These values correspond to relative error reductions of 45.9% and 57.1%, respectively. Under moderate occlusion (25 to 70%), the MAE reaches 3.8 bpm, while PhysFormer and MTTS-CAN yield 8.7 bpm and 10.2 bpm, respectively, resulting in reductions of 56.3% and 62.7%. Similar trends are observed in light and minimal occlusion conditions. These findings underscore the ability of our adaptive masking and distillation strategy to maintain signal integrity in the presence of visual obstructions.

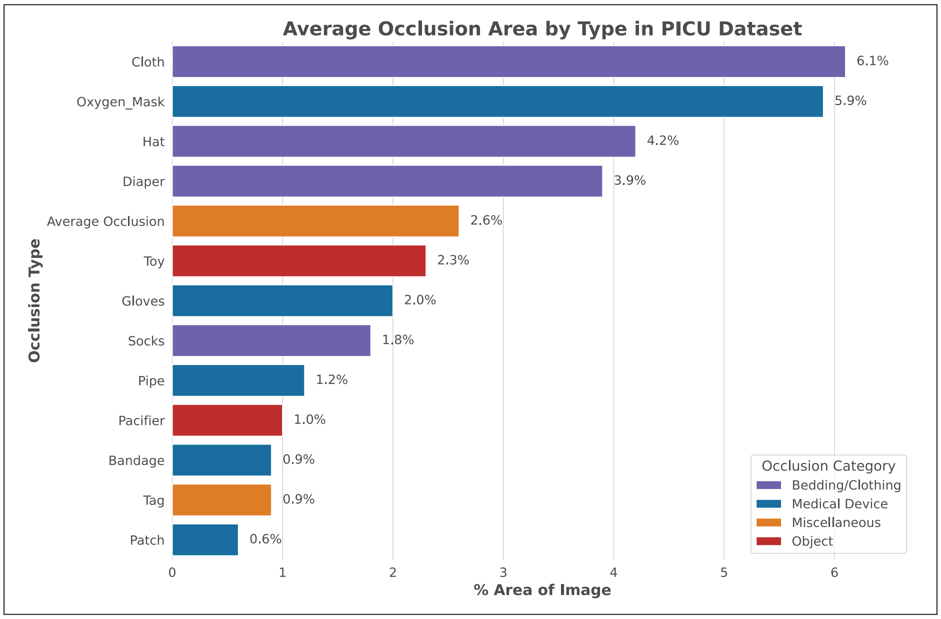


Figure 4.5 Average occlusion area by type in the PICU dataset. Medical devices and bedding/clothing are the most common and largest sources of occlusion

Table 4.3 Performance Under Different Occlusion Levels

Occlusion Level	**Our Method** MAE (bpm)	**PhysFormer** MAE (bpm)	**MTTS-CAN** MAE (bpm)
None (<10%)	2.3	4.8	5.9
Light (10–25%)	2.9	6.1	7.4
Moderate (25–70%)	3.8	8.7	10.2
Severe (>70%)	7.2	13.3	16.8

4.6.5 Cross-dataset Evaluation on UBFC with Synthetic Occlusions

We evaluated the staged pretraining strategy on the UBFC-rPPG dataset under controlled synthetic occlusions to examine the model's ability to generalize across domains. As presented

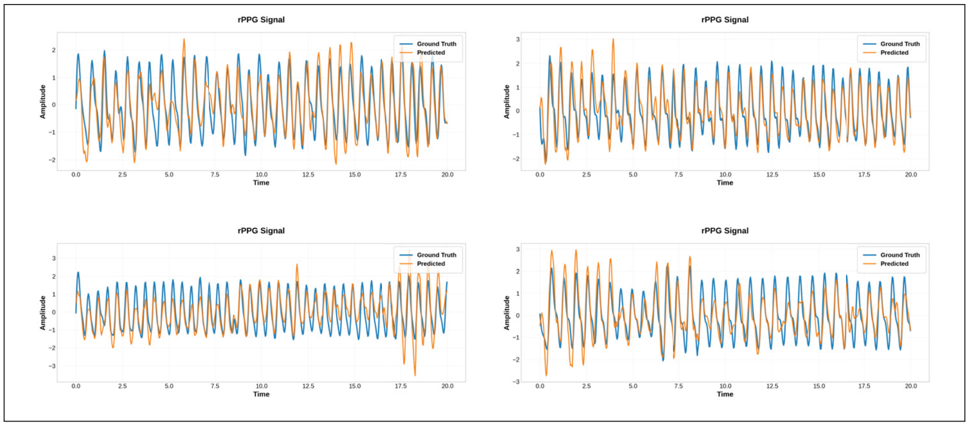


Figure 4.6 Time-domain overlay of predicted and ground truth rPPG signals for four UBFC subjects with synthetic occlusions

in Table 4.4, the heart rate estimation error progressively decreases over the three stages (MAE: 5.4 → 3.1 → 2.8 bpm). The addition of Stage 2, which introduces occlusion augmentation, results in a significant reduction in error, suggesting that the training pipeline enhances robustness to obstructed facial regions and improves generalization to clinical-like conditions.

Figure 4.6 displays examples from four UBFC subjects with 40 to 60% facial occlusion, including simulated masks, tubing, and overlays. Despite reduced visibility, the predicted rPPG signals remain temporally consistent and closely follow the ground truth. These qualitative observations reinforce the quantitative findings and demonstrate that curriculum-guided self-supervised learning improves model reliability under occlusion and domain shift.

Table 4.4 UBFC-Occluded results (20 s windows)

Training Stage	**MAE**	**RMSE**	**R**
Stage 1 only	5.4	13.0	0.76
Stage 1+2	3.1	4.7	0.89
Full Pipeline	**2.8**	**4.1**	**0.93**

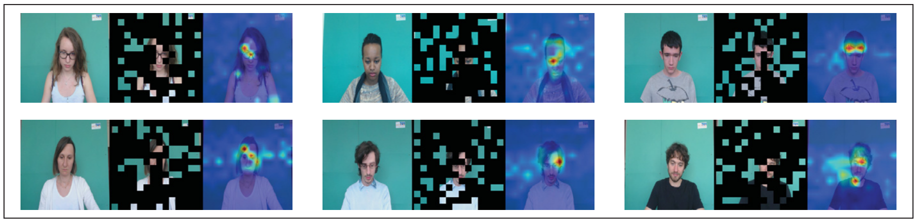


Figure 4.7 Grad-CAM visualizations illustrating the spatial attention of our student model on six UBFC subjects. Each triplet includes: the original RGB input, the masked version with 75% token dropout from the AMN, and the corresponding Grad-CAM heatmap from the final rPPG decoder layer. Despite adversarial masking, the model consistently attends to physiologically relevant facial areas, primarily the forehead and upper cheeks, without relying on predefined regions of interest. These results confirm the model's ability to learn spatial priors required for accurate rPPG estimation, even under partial occlusion

4.6.6 Qualitative Analysis and Model Interpretability

Figure 4.7 presents Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations from six UBFC subjects to examine the spatial attention learned by the student model. For each subject, three views are shown: the original RGB frame, the masked input with 75% token dropout from the Adaptive Masking Network (AMN), and the Grad-CAM activation map extracted from the final rPPG decoder layer.

Despite strong masking during pretraining, the model consistently focuses on physiologically relevant facial regions, particularly the forehead, periorbital areas, and upper cheeks. These regions correspond to areas with dense superficial vasculature, such as the temporal and facial arteries, which are optimal for pulse extraction. This spatial selectivity emerges without explicit supervision or predefined regions of interest, indicating that the combination of reconstruction objectives and distillation effectively encodes physiological priors.

The masking strategy contributes significantly to this selective attention behavior. By intentionally suppressing informative tokens during pretraining, the AMN forces the model to rely on the remaining visible facial regions for pulse recovery. Grad-CAM analysis shows that the model adapts dynamically, shifting its focus toward alternative vascular regions when primary areas are occluded, while maintaining physiological consistency. This adaptive mechanism directly supports the quantitative findings in Table I-4, where the model retains a MAE of 5.2 bpm under more than 50% occlusion, compared to 14.3 bpm for PhysFormer.

The visualizations also reveal strong spatial selectivity, with minimal activation in background and masked regions. The model concentrates attention on exposed skin areas, independent of skin tone, facial structure, or the presence of accessories such as glasses. This suggests that attention is guided by physiological cues rather than superficial appearance.

These qualitative observations support the hypothesis that combining adaptive masking with physiological distillation enables the model to identify clinically relevant features without handcrafted priors. The model also maintains accurate pulse estimation under severe occlusions. This performance, achieved through self-supervised pretraining followed by limited supervised fine-tuning, confirms the ability of the framework to learn robust and physiologically grounded representations.

To evaluate the generalization of the learned attention behavior, we further analyzed Grad-CAM visualizations from real PICU patients recorded under diverse conditions. Figure 4.8 shows examples from 14 representative cases with varying levels of occlusion, skin tone, and the presence of medical equipment.

The model autonomously identifies physiologically meaningful areas, including the forehead, periorbital regions, upper thorax, and cheeks, without relying on predefined facial landmarks or anatomical masks. These are the regions where superficial blood vessels are most accessible, and the model consistently attends to them even when part of the face is covered.

The attention maps indicate adaptive spatial behavior in the presence of clinical occlusions such as oxygen masks or medical tubing. When primary pulse-bearing regions are blocked, the model redistributes its attention toward secondary areas that preserve pulsatile information, maintaining signal continuity. This adaptive mechanism shows that the model prioritizes physiological relevance rather than fixed spatial locations.

The learned attention strategy remains consistent across patient age groups and clinical conditions. Similar spatial focus is observed in neonates (e.g., P6) and older children (e.g., P3, P11), as well as across skin tones (e.g., P1, P4 vs. P9, P12). This invariance supports the physiological basis of the attention mechanism and suggests that region selection is driven by underlying pulse characteristics.

Compared to the laboratory results in Figure 4.7, the PICU attention maps show sharper spatial localization and more targeted region prioritization. This improvement likely results from adversarial masking during pretraining, which encourages reliance on the most informative pulse-bearing regions. When these areas are occluded, the model broadens its focus to secondary visible regions, indicating a compensatory mechanism that maintains signal integrity.

This automatic discovery of relevant regions removes the need for separate region-of-interest detection steps, reducing preprocessing complexity and improving robustness. The model learns to segment and weight physiologically meaningful areas end-to-end, enabling practical deployment in clinical monitoring environments.

4.6.7 Progressive Pretraining Stages

Table 4.5 Ablation of Progressive Learning Stages

Training Configuration	**MAE** (bpm)	**RMSE** (bpm)	**R**	**ΔBland-Altman** (95% CI)
Direct Supervised (160 patients)	18.2	22.4	0.41	-35.2 to 38.6
Stage 1 only → Fine-tune	10.3	13.6	0.64	-21.3 to 22.8
Stage 1+2 → Fine-tune	6.8	9.2	0.78	-14.2 to 15.7
Stage 1+2+3 → Fine-tune (Full)	**3.2**	**5.4**	**0.91**	**-6.8 to 7.2**

Table 4.5 quantifies the cumulative benefits of our three-stage curriculum learning strategy. Compared to direct supervised training on 160 PICU patients, each stage contributes measurable performance improvements that accumulate across the pipeline.

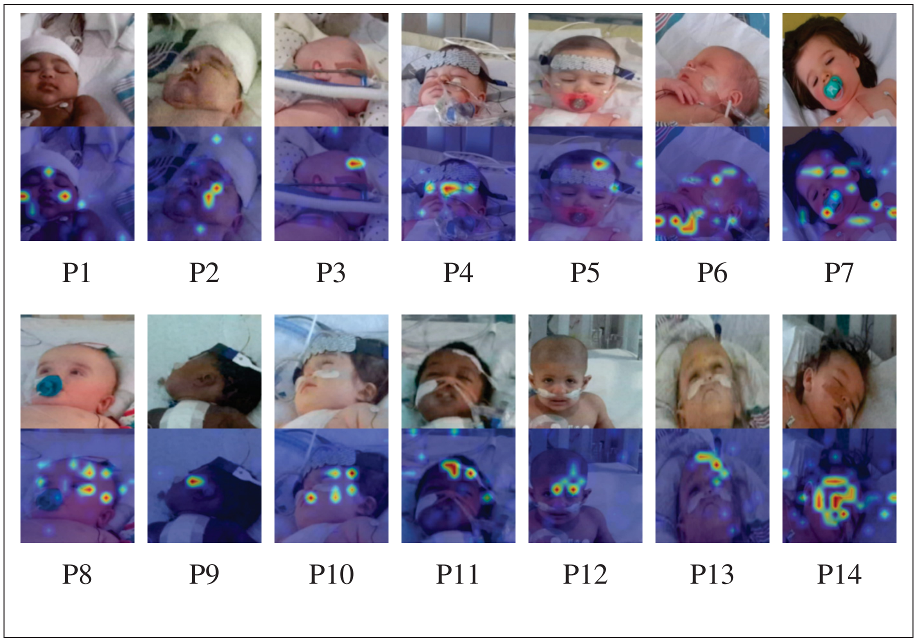


Figure 4.8 Grad-CAM visualizations from 14 PICU patients. Each pair shows the original RGB frame (top) and the corresponding Grad-CAM heatmap (bottom) generated from the final rPPG decoder layer. Attention consistently focuses on physiologically informative areas such as the forehead, periorbital region, upper cheeks, and thorax, while background regions and medical equipment receive minimal activation. These results confirm the model's ability to adapt spatial attention across varying occlusion levels, skin tones, and patient conditions, without relying on predefined anatomical priors. All parents gave their written consent to present their child image in a scientific publication

Stage 1 establishes core physiological representations using self-supervised pretraining on clean public datasets. This alone reduces MAE from 18.2 bpm to 10.3 bpm and raises the correlation from 0.41 to 0.64, confirming that exposure to controlled conditions enhances the model's ability to extract generalizable pulse features.

Stage 2 introduces synthetic clinical artifacts through our hospital occlusion simulator, preparing the model for real-world challenges. Critically, these synthetic occlusions force the model to search beyond easily accessible regions and discover alternative physiological pathways for pulse detection. By dynamically masking different facial regions during training, the model is compelled to explore multiple pathways for rPPG estimation rather than relying on fixed spatial cues. This encourages learning signal-rich zones based on their physiological utility, not on explicit ROI supervision. Consequently, the model becomes more robust to occlusions and better aligned with the anatomical variability encountered in clinical videos. As a result, MAE further decreases to 6.8 bpm and correlation rises to 0.78.

Stage 3 bridges the domain gap by adapting the model to unlabeled PICU videos. Fine-tuning on this real-world data achieves a final MAE of 3.2 bpm and a correlation of 0.91. The complete pipeline yields significant gains in both accuracy and robustness by allowing the model to learn spatial priors from data rather than from explicit supervision.

Overall, the results validate our hypothesis: progressive complexity scheduling, coupled with spatially unconstrained learning, enables effective adaptation to noisy, occluded, and clinically diverse conditions, even under limited ground truth supervision.

4.6.8 Adaptive Masking Strategy

Figure 4.9 presents the temporal evolution of our adaptive masking strategy over 128 frames. The middle row shows that the AMN consistently assigns high importance scores to key pulse-bearing regions such as the forehead and upper cheeks. Despite this spatial focus, the bottom row demonstrates that the AMN generates temporally diverse masking patterns while preserving the target 25% visibility ratio. This behavior prevents the student from memorizing fixed occlusion patterns and encourages the extraction of robust, redundant features from various facial regions.

Figure 4.10 quantifies the statistical properties of the learned masking policy. The predicted importance scores follow a bimodal distribution, with distinct peaks at approximately 0.20 and 0.75, corresponding to non-informative and physiologically relevant regions, respectively. The spatial average map exhibits a strong center bias with exponential attenuation toward the periphery, highlighting the AMN's spatial selectivity. Temporally, the patch visibility remains stable around 25% ±2%, confirming that the model does not collapse into fixed or static configurations. These

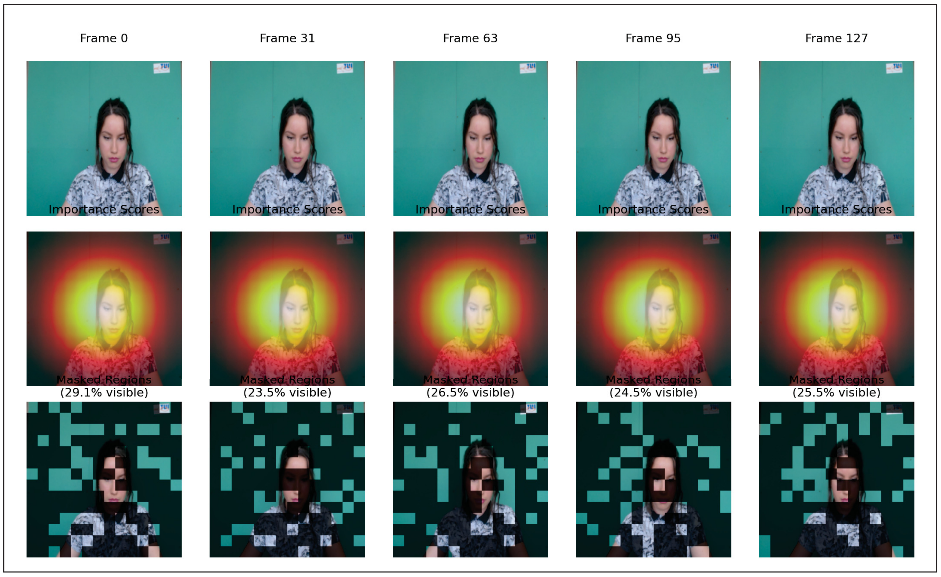


Figure 4.9 Temporal progression of adaptive masking across 128 frames. Top: selected input frames at $t = 0, 31, 63, 95, 127$. Middle: AMN-generated importance maps consistently emphasize physiologically salient regions (e.g., forehead, upper cheeks). Bottom: resulting masked inputs with ~25% visibility, demonstrating spatial consistency and temporal diversity in visible patch selection

results suggest that the AMN has converged to a stable yet dynamic masking policy that balances spatial selectivity with temporal variation. By enforcing reconstruction from different patch subsets, the model is pushed to learn redundant, generalizable representations of pulse-rich facial regions, which directly supports robustness under real-world occlusion.

Table 4.6 evaluates the impact of different masking strategies. Our full AMN approach achieves the lowest MAE (3.2 bpm) and highest correlation (0.91), significantly outperforming baseline methods. Compared to tube-based masking (5.2 bpm), adaptive masking yields a 38.5% improvement in MAE. Random masking strategies are less effective, and removing the policy gradient degrades performance to 4.8 bpm. Although the full AMN increases training time to 1.8, its gains in accuracy and robustness justify the computational cost for clinical deployment. These results support the utility of adversarial masking to promote learning of transferable features across occluded and unconstrained facial inputs.

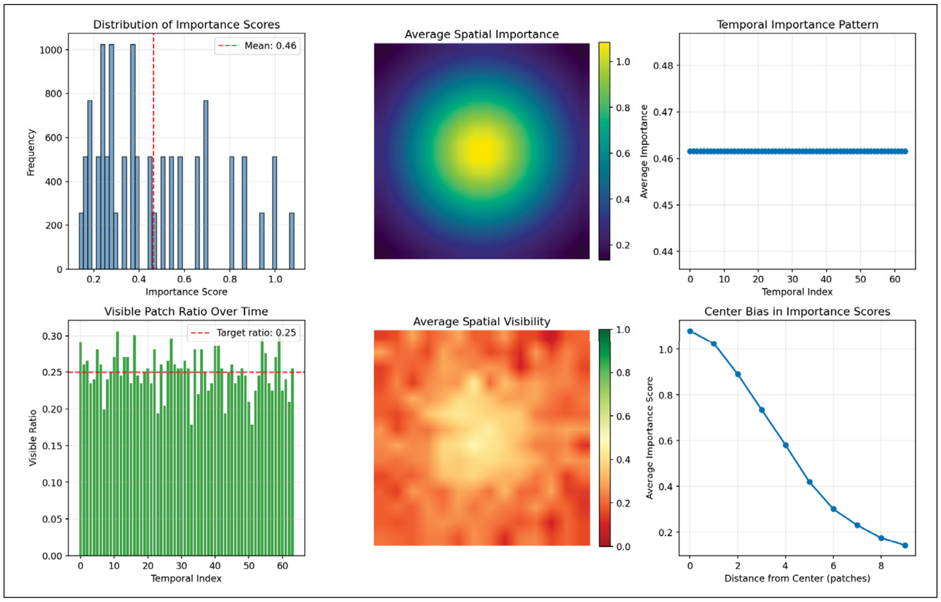


Figure 4.10 Quantitative analysis of the learned masking policy across 128 frames. Importance score distribution shows clear bimodality, separating background patches (~0.20) from facial regions (~0.75). The average spatial map reveals strong central activation. The radial profile exhibits exponential decay from the center to periphery (7.3× contrast), reflecting spatial selectivity

Table 4.6 Ablation of Masking Strategies

Masking Strategy	MAE (bpm)	RMSE (bpm)	R	Training Time
No Masking (Standard AE)	8.4	11.2	0.70	1.0x
Random Masking (75%)	5.6	7.9	0.83	1.2x
Random Masking (90%)	6.1	8.4	0.80	1.2x
Tube Masking (VideoMAE)	5.2	7.5	0.85	1.3x
Adaptive (w/o Policy Gradient)	4.8	7.1	0.86	1.5x
Adaptive (Full AMN)	**3.2**	**5.4**	**0.91**	**1.8x**

4.6.9 Loss Function Components

Table 4.7 summarizes the impact of each loss component and their combinations. Adding the correlation loss to the pixel-wise MSE reduces the MAE from 7.3 to 5.1 bpm, corresponding to a 30.1% improvement. When used alone, the distillation objective achieves an MAE of 4.6 bpm, representing a 37.0% reduction compared to MSE. The combination of both objectives yields the lowest MAE of 3.2 bpm, resulting in a 56.2% improvement over MSE, 37.3% over MSE+Corr, and 30.4% over MSE+Distill. These results confirm that pixel-level fidelity and physiological guidance act as complementary supervisory signals, improving both estimation accuracy and temporal consistency.

Table 4.7 Ablation of Loss Components

Loss Configuration	**MAE (bpm)**	**RMSE (bpm)**	**R**
MSE only (L_{pixel})	7.3	9.8	0.73
MSE + Correlation (L_{recon})	5.1	7.3	0.84
MSE + Distillation	4.6	6.8	0.87
Full ($L_{\textbf{recon}}$ + $L_{\textbf{distill}}$)	**3.2**	**5.4**	**0.91**

4.6.10 Architecture Components

In critical care environments such as the PICU, real-time inference and hardware constraints demand efficient and deployable models. VisionMamba was selected for its low-latency and lightweight design, offering a favorable trade-off between speed, memory, and accuracy. Table 4.8 presents a comparative analysis of computational metrics across several backbone architectures.

VisionMamba achieves the fastest inference time at 85 ms, outperforming ViT-B by 7 ms and reducing latency by 17%. It also requires the fewest operations (15.5 GFLOPs) and consumes the least memory (1.1 GB), while maintaining a compact model size of 50.2M parameters. These characteristics support its integration into clinical pipelines, where uninterrupted, low-overhead processing is essential for continuous rPPG estimation without interfering with patient care.

Table 4.8 Comparison of computational efficiency between encoder backbones

Architecture	**Params (M)**	**FLOPs (G)**	**Memory (GB)**	**Inference Time (ms)**
ViT-B (Baseline)	86.4	16.1	7.3	92
TimeSformer	121.3	379.8	8.6	300
VideoMAE	86.2	101.85	7.9	108
VisionMamba (Ours)	**50.2**	**15.5**	**1.1**	**85**

4.7 Conclusion

Continuous and unobtrusive monitoring of vital signs in the PICU requires models that are accurate, lightweight, and robust to real-world clinical conditions. This work presents a self-supervised rPPG estimation framework tailored to these demands. Based on a VisionMamba backbone, the proposed system integrates an Adaptive Masking Network to enforce spatial-temporal relevance and a physiological distillation head to guide signal consistency. Evaluated on realistic PICU videos, the method achieves 3.2 bpm MAE with strong temporal correlation, while operating within clinical latency and resource constraints (85 ms inference time, 15.5 GFLOPs, 1.1 GB memory). The masking strategy provides spatial selectivity and temporal diversity under occlusion, and the distillation mechanism enhances physiological fidelity across sequences. These design choices allow the model to generalize across challenging lighting conditions and partial obstructions common in bedside recordings. Future work will focus on expanding the demographic diversity of the training data, conducting multi-institutional validation, and extending the approach to respiration monitoring. Additional efforts will address real-time adaptation and integration with embedded systems to enable secure, continuous inference in the clinical workflow.

CONCLUSION AND RECOMMENDATIONS

This thesis addressed the urgent need for robust, non-contact physiological monitoring in pediatric intensive care environments, where traditional sensor-based approaches remain invasive, error-prone, and often infeasible for neonates and critically ill infants. By developing and evaluating a unified framework combining efficient deep learning architectures, curriculum-based self-supervised learning, and real-time anatomical detection, this work advances the field of medical video analysis toward practical clinical deployment.

To bridge the gap between laboratory prototypes and clinical deployment in the PICU, this thesis tested the overarching hypothesis that physiological signal integrity in complex environments relies on the synergy between temporal gradient modeling, robust anatomical tracking, and curriculum-based domain adaptation. While traditional approaches treat rPPG as a pure signal processing task, our findings demonstrate that it must be treated as a holistic computer vision challenge. We validated this through three interconnected methodological advancements.

First, regarding pulsatile dynamics, we demonstrated that standard 3D-CNNs fail to capture the subtle, quasi-periodic nature of the pulse. By introducing convolutional networks with 3D Temporal Central Difference Convolutions (3DCDC-T) combined with global self- attention, we confirmed that explicitly modeling temporal gradients, rather than absolute intensity, is required to isolate the Blood Volume Pulse (BVP) from rigid head motions. This model achieved high accuracy in heart rate estimation, outperforming state-of-the-art methods across public and clinical datasets.

Second, regarding temporal stability, we addressed the hypothesis that frame-independent detectors introduce fatal jitter in physiological signals. By developing DST-Mamba, we proved that decoupling spatial and temporal state-space modeling achieves linear-time complexity while maintaining the orientation-aware tracking necessary for vital signs estimation. This confirms that detection in the PICU requires memory of past states, not just spatial recognition.Its use of oriented bounding boxes improved motion-aligned cropping, and RGB-D fusion enhanced resilience under partial occlusion, achieving stable tracking with minimal latency. The model demonstrated strong localization performance, achieving 0.96 mAP@0.5 and 0.95 rotated IoU.

Third, regarding data scarcity, we validated that a curriculum-based self-supervised learning approach could effectively transfer features from clean domains to the noisy PICU environment. By utilizing an Adaptive Masking Network, we demonstrated that adversarially hiding informative regions forces the model to learn robust, redundant physiological representations, reducing the reliance on labeled data while preserving clinically acceptable accuracy.

This framework establishes that efficient, domain-adapted deep learning models can reliably perform vital sign monitoring from video alone, supporting both signal extraction and anatomical tracking in complex ICU settings. The use of self-supervised pretraining and efficient sequence modeling narrows the gap between research prototypes and deployable bedside systems.

The proposed system has not yet undergone prospective evaluation in real-time hospital workflows. We recommend initiating clinical trials across multiple PICU environments to validate performance in diverse lighting, camera configurations, and patient populations. The current dataset, while extensive, is institution-specific. Generalization to other settings will require standardized calibration protocols for depth sensors and consistent positioning of camera equipment to ensure reproducibility of results (Yang, Soltan & Clifton, 2022a; Hornback *et al.*, 2022).

Integration into clinical infrastructure should align with privacy-preserving edge AI paradigms, allowing inference and model adaptation to occur on-device (Rieke *et al.*, 2020). This approach minimizes latency and respects healthcare data regulations. Clinicians should be provided with confidence scores and alert flags to guide interpretation, particularly in scenarios involving high occlusion or motion artifacts. The system should be considered a supplementary monitoring tool and not a replacement for contact-based sensors until further validation is conducted.

Training for ICU staff will be necessary to ensure correct usage, particularly regarding the model's limitations in detecting signals under severe occlusion or in the presence of movement artifacts (Knop, Weber, Mueller & Niehaves, 2022). Clear protocols must be defined for transitioning between contact-based and non-contact modalities based on detection confidence and clinical conditions.

Several directions can further extend the impact of this work. First, future efforts should integrate the detection and rPPG estimation components into a unified, end-to-end pipeline. Such integration would allow for signal quality to guide region refinement in real time, improving robustness in unstable scenarios. Second, extending the framework to include additional modalities, such as respiratory motion, oxygen saturation, and body temperature, would support multimodal physiological monitoring. This expansion requires the alignment and synchronization of heterogeneous data streams for fusion at both the spatial and temporal levels (Hassanpour & Yang, 2025).

Third, deploying the model on embedded clinical systems such as bedside monitors or edge devices will require compression techniques including model pruning, quantization, and architectural simplification (Zhu & Gupta, 2017). The current GPU-based performance is promising, but inference under CPU-only constraints remains a critical step for hospital integration.

Fourth, the curriculum-based masking strategy, while effective in boosting robustness, requires further investigation to optimize the sequence of tasks and synthetic perturbations. Realistic data augmentation that preserves physiological signal integrity should be explored to avoid introducing biases. Fifth, demographic fairness and performance equity must be rigorously evaluated. Differences in skin tone, age, or facial development can impact signal visibility and detection performance. Stratified validation and fairness-aware learning objectives should be adopted to ensure equitable care.

Future studies should explore continuous angle regression techniques to improve the orientation precision of detected bounding boxes, avoiding the discretization artifacts inherent to the CSL representation. Temporal stability can be further enhanced by integrating recurrent memory modules or explicit consistency constraints into the loss function. Additionally, the use of large-scale synthetic datasets for pretraining, while valuable, should be augmented with real-world variations in motion, lighting, and occlusion to improve generalization.

More advanced fusion strategies, such as cross-attention or graph-based modeling of facial subregions, may better capture physiological interdependencies between regions. Finally, transformer patch sizes and embedding dimensions, currently inherited from vision transformer baselines, should be optimized for the specific resolution and content of PICU video streams, with a focus on preserving clinically relevant spatiotemporal information.

The techniques developed in this thesis have applicability beyond intensive care. Non-contact monitoring may benefit outpatient settings, neonatal follow-up programs, and remote home care, particularly for immunocompromised or preterm infants. However, ethical deployment will require careful attention to data privacy, algorithm transparency, and clinical oversight.

Although the source code and augmentation framework will be made available to the research community, access to clinical video data remains restricted. Collaborative frameworks that allow federated benchmarking or synthetic patient simulations will be essential to expand impact while respecting patient confidentiality.

This work contributes to the growing body of research demonstrating that intelligent, efficient, and self-supervised architectures can meet the unique constraints of pediatric critical care. The integration of robust detection, signal estimation, and domain-specific learning offers a pathway to transform how vital signs are monitored in the ICU, aligning machine intelligence with real-world clinical needs.

APPENDIX I

FIGURES AND TABLES - CHAPTER 3

1. Figures

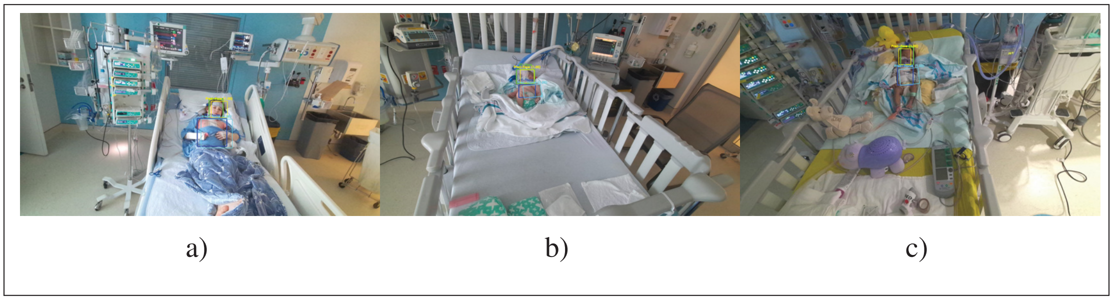


Figure-A I-1 Representative failure cases under challenging conditions. (a) partial patient occlusion of the face or thoracoabdominal region by medical equipment or coverings. (b) severe occlusion from medical devices. (c) low-illumination scene.

2. Tables

Table-A I-1 Ablation study on model components. This table shows the impact of removing or adding specific features; such as angle loss, fixed orientation, rotated IoU (rIoU), and depth information; on overall detection performance.

Model Variant	**IoU**	**mAP@0.50**	**Angle**
Without Angle Loss	0.81	0.33	0.37
Without Orientation	0.92	0.487	0.00
Without rIoU	0.90	0.92	0.40
With Depth	**0.96**	**0.95**	0.52

Table-A I-2 Impact of masking ratio on self-supervised pre-training efficiency and downstream task performance.

Masking Ratio	Fine-tune Epochs	mAP@0.5	Memory (GB)
50%	95	0.89	12.28
60%	85	0.91	10.73
70%	80	0.93	9.22
80%	**70**	**0.95**	**7.73**
90%	70	0.94	6.21
95%	75	0.92	5.45

Table-A I-3 Impact of Mamba encoder depth on model performance and efficiency. All experiments use 224×224 input resolution with 16-frame clips. The optimal configuration is highlighted.

Layers	Parameters (M)	FLOPs (G)	mAP@0.50	rIoU
4	43.2	0.85	0.88	0.90
8	53.1	1.15	0.92	0.93
12	**73.7**	**1.85**	**0.95**	**0.96**
16	96.4	2.45	0.96	0.97

Table-A I-4 Performance Across Occlusion Levels.

Occlusion Level	IoU	mAP@0.50	mAP50-95
None	0.96	0.98	0.68
Light	0.95	0.96	0.61
Moderate	0.88	0.89	0.52
Severe	0.61	0.58	0.31

BIBLIOGRAPHY

Abusamra, H. N. J., Ali, S. H. M., Elhussien, W. A. K., Mirghani, A. M. A., Ahmed, A. A. A. & Ibrahim, M. E. A. (2025). Ethical and Practical Considerations of Artificial Intelligence in Pediatric Medicine: A Systematic Review. *Cureus*, 17(2), e79024. doi: 10.7759/cureus.79024.

Al-Sofyani, K. A. (2025). Role of Artificial Intelligence in Pediatric Intensive Care: A Survey of Healthcare Staff Perspectives in Saudi Arabia. *Frontiers in Pediatrics*, 13, 1533877. doi: 10.3389/fped.2025.1533877.

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3), R1. doi: 10.1088/0967-3334/28/3/R01.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M. & Schmid, C. (2021). ViViT: A Video Vision Transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1(1), 6836–6846.

Association for the Advancement of Medical Instrumentation. (2002). ANSI/AAMI EC13:2002 – Cardiac Monitors, Heart Rate Meters, and Alarms. Retrieved from: https://webstore.ansi.org/standards/aami/ansiaamiec132002.

Bao, H., Dong, L., Piao, S. & Wei, F. (2021). BEiT: BERT Pre-Training of Image Transformers. *arXiv preprint arXiv:2106.08254*, abs/2106.08254(1), 1–10.

Bautista, C., Gan, J. H., Tan, K. H., Gan, C. H., Ling, S., Chan, M. Y., Ong, L. Y. & Sim, A. S. C. (2023). Clinical Applications of Contactless Photoplethysmography for Vital Signs Monitoring in Paediatrics: A Systematic Review and Meta-Analysis. *Journal of Clinical Monitoring and Computing*, 37(4), 915–928. doi: 10.1007/s10877-023-01045-8.

Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K. & Grundmann, M. (2019). BlazeFace: Sub-Millisecond Neural Face Detection on Mobile GPUs. *arXiv preprint arXiv:1907.05047*, abs/1907.05047(1), 1–9.

Behrouz, A. & Hashemi, F. (2024). Graph mamba: Towards learning on graphs with state space models. *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 119–130.

Benjamens, S., Dhunnoo, P. & Meskó, B. (2020). The State of Artificial Intelligence-Based FDA-Approved Medical Devices and Algorithms: An Online Database. *NPJ Digital Medicine*, 3(1), 118. doi: 10.1038/s41746-020-00324-0.

Bertasius, G., Wang, H. & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? *Proceedings of the International Conference on Machine Learning (ICML)*, 2(3), 4.

Bewley, A., Ge, Z., Ott, L., Ramos, F. & Upcroft, B. (2016). Simple Online and Realtime Tracking. *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, 1(1), 3464–3468.

Bhattachrjee, S., Li, H., Xia, J. & Xu, W. (2023). SimPPG: Self-supervised photoplethysmography-based heart-rate estimation via similarity-enhanced instance discrimination. *Smart Health*, 28, 100396.

Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A. & Dubois, J. (2019). Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124, 82–90.

Boivin, V., Shahriari, M., Faure, G., Mellul, S., Tiassou, E. D., Jouvet, P. & Noumeir, R. (2023). Multimodality Video Acquisition System for the Assessment of Vital Distress in Children. *Sensors*, 23(11), 5293.

Bonafide, C. P., Lin, R., Zander, M., Graham, C. S., Paine, C. W., Rock, W. & Keren, R. (2015). Association Between Exposure to Nonactionable Physiologic Monitor Alarms and Response Time in a Children's Hospital. *Journal of Hospital Medicine*, 10(6), 345–351. doi: 10.1002/jhm.2331.

Bondarenko, M., Menon, C. & Elgendi, M. (2025). Demographic Bias in Public Remote Photoplethysmography Datasets. *NPJ Digital Medicine*, 8, 593. doi: 10.1038/s41746-025-01973-9.

Bostan, I., van Egmond, R., Gommers, D. & Özcan, E. (2024). Customizing ICU Patient Monitoring: A User-Centered Approach Informed by Nurse Profiles. *Cognition, Technology & Work*, 26(3), 507–522. doi: 10.1007/s10111-023-00751-5.

Botina-Monsalve, D., Benezeth, Y. & Miteran, J. (2022). RTrPPG: An Ultra Light 3DCNN for Real-Time Remote Photoplethysmography. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2146–2154.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11), 120–123.

Cao, M., Cheng, X., Liu, X., Jiang, Y., Yu, H. & Shi, J. (2024). St-phys: Unsupervised spatio-temporal contrastive remote physiological measurement. *IEEE Journal of Biomedical and Health Informatics*, 28(8), 4613–4624.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. *Proceedings of the European Conference on Computer Vision (ECCV)*, 12346(1), 213–229.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660.

Carreira, J. & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1(1), 6299–6308.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D. & Sutskever, I. (2020a). Generative pretraining from pixels. *International conference on machine learning*, pp. 1691–1703.

Chen, S., Ma, K. & Zheng, Y. (2019). Med3D: Transfer Learning for 3D Medical Image Analysis. *arXiv preprint arXiv:1904.00625*, abs/1904.00625(1), 1–9.

Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. (2020b). A simple framework for contrastive learning of visual representations. *International conference on machine learning*, pp. 1597–1607.

Chen, W. & McDuff, D. (2018). DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. In Ferrari, V., Hebert, M., Sminchisescu, C. & Weiss, Y. (Eds.), *Computer Vision – ECCV 2018* (vol. 11206, pp. 356–373). Cham: Springer. doi: 10.1007/978-3-030-01216-8_22.

Cheng, C.-H., Wong, K.-L., Chin, J.-W., Chan, T.-T. & So, R. H. Y. (2021). Deep Learning Methods for Remote Heart Rate Measurement: A Review and Future Research Agenda. *Sensors*, 21(18), 6296.

Cvach, M. (2012). Monitor Alarm Fatigue: An Integrative Review. *Biomedical Instrumentation & Technology*, 46(4), 268–277. doi: 10.2345/0899-8205-46.4.268.

Dai, Z., Liu, H., Le, Q. V. & Tan, M. (2021). CoAtNet: Marrying Convolution and Attention for All Data Sizes. *Advances in Neural Information Processing Systems*, 34, 3965–3977.

Dasari, A., Prakash, S. K. A., Jeni, L. A. & Tucker, C. S. (2021). Evaluation of biases in remote photoplethysmography methods. *NPJ digital medicine*, 4(1), 91.

De Haan, G. & Jeanne, V. (2013). Robust Pulse Rate From Chrominance-Based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10), 2878–2886. doi: 10.1109/TBME.2013.2266196.

Deng, J., Guo, J., Ververas, E., Kotsia, I. & Zafeiriou, S. (2020). RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1(1), 5203–5212.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv preprint arXiv:2010.11929. Retrieved from: https://arxiv.org/abs/2010.11929.

Dosso, Y. S., Kyrollos, D., Greenwood, K. J., Harrold, J. & Green, J. R. (2022). NICUface: Robust Neonatal Face Detection in Complex NICU Scenes. *IEEE Access*, 10(1), 62893–62909.

Eastwood-Sutherland, C., Lim, K., Gale, T. J., Wheeler, K. I. & Dargaville, P. A. (2023). Detection of Respiratory Activity in Newborn Infants Using a Noncontact Vision-Based Monitor. *Pediatric Pulmonology*, 58(6), 1753–1760.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2), 303–338.

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J. & Feichtenhofer, C. (2021). Multiscale Vision Transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1(1), 6824–6835.

Fayyaz, M., Bahrami, E., Diba, A., Noroozi, M., Adeli, E., Van Gool, L. & Gall, J. (2021). 3D CNNs with adaptive temporal feature resolutions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4731–4740.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J. & Saria, S. (2021). The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine*, 385(3), 283–286. doi: 10.1056/NEJMc2104626.

Fu, D. Y., Dao, T., Saab, K. K., Thomas, A. W., Rudra, A. & Ré, C. (2022). Hungry Hungry Hippos: Towards Language Modeling with State Space Models. *arXiv preprint arXiv:2212.14052*, abs/2212.14052(1), 1–12.

Gideon, J. & Stent, S. (2021). The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3995–4004.

Girshick, R. (2015). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1(1), 1440–1448.

Girshick, R., Donahue, J., Darrell, T. & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1(1), 580–587.

Google Developers. (2024). Module: MediaPipe. *Google for Developers*, 1(1), 1–1. [Online]. Available: https://developers.google.com/mediapipe/api/solutions/python/mp.

Görges, M., Kück, K., Koch, S. H., Agutter, J. & Westenskow, D. R. (2011). A Far-View Intensive Care Unit Monitoring Display Enables Faster Triage. *Dimensions of Critical Care Nursing*, 30(4), 206–217. doi: 10.1097/DCC.0b013e31821a0586.

Graham, K. C. & Cvach, M. (2010). Monitor Alarm Fatigue: Standardizing Use of Physiological Monitors and Decreasing Nuisance Alarms. *American Journal of Critical Care*, 19(1), 28–34. doi: 10.4037/ajcc2010651.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M. et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33, 21271–21284.

Grooby, E., Sitaula, C., Ahani, S., Holsti, L., Malhotra, A., Dumont, G. A. & Marzbanrad, F. (2023a). Neonatal Face and Facial Landmark Detection from Video Recordings. *Proceedings of the 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1(1), 1–5.

Grooby, E., Sitaula, C., Ahani, S., Holsti, L., Malhotra, A., Dumont, G. A. & Marzbanrad, F. (2023b). Neonatal Face and Facial Landmark Detection from Video Recordings. *Proceedings of the 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1(1), 1–5.

Gu, A. & Dao, T. (2023). *Mamba: Linear-Time Sequence Modeling With Selective State Spaces*. arXiv preprint arXiv:2312.00752. Retrieved from: https://arxiv.org/abs/2312.00752.

Gu, A., Goel, K. & Ré, C. (2021). Efficiently Modeling Long Sequences with Structured State Spaces. *arXiv preprint arXiv:2111.00396*, abs/2111.00396(1), 1–10.

Guan, H. & Liu, M. (2021). Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Transactions on Biomedical Engineering*, 69(3), 1173–1185.

Gudi, A., Bittner, M. & van Gemert, J. (2020). Real-Time Webcam Heart-Rate and Variability Estimation with Clean Ground Truth for Evaluation. *Applied Sciences*, 10(23), 8630.

Hasan, Z., Faridee, A. Z. M., Ahmed, M. & Roy, N. (2022). Self-rppg: Learning the optical & physiological mechanics of remote photoplethysmography with self-supervision. *2022 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 46–56.

Hassanpour, A. & Yang, B. (2025). Contactless Vital Sign Monitoring: A Review Towards Multi-Modal Multi-Task Approaches. *Sensors*, 25(15), 4792.

Hatib, A. N. A., Lee, J., Chong, S. L., Ang, Y., Choo, H. H., Chong, S.-L. & Tham, E. H. (2024). A two-phased study on the use of remote photoplethysmography (rPPG) in paediatric care. *Biomedical Signal Processing and Control*, 89, 105050. doi: 10.1016/j.bspc.2023.105050.

Hausmann, J., Salekin, M. S., Zamzmi, G., Goldgof, D. & Sun, Y. (2022). Robust Neonatal Face Detection in Real-World Clinical Settings. *arXiv preprint arXiv:2204.00655*, abs/2204.00655(1), 1–8.

He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. & Girshick, R. (2022). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009.

Hedegaard, L. & Iosifidis, A. (2022). Continual 3D convolutional neural networks for real-time processing of videos. *European Conference on Computer Vision*, pp. 369–385.

Helman, S., Terry, M. A., Pellathy, T., Williams, A., Dubrawski, A., Clermont, G. & Hravnak, M. (2022). Engaging Clinicians Early During the Development of a Graphical User Display of an Intelligent Alerting System at the Bedside. *International Journal of Medical Informatics*, 159, 104643. doi: 10.1016/j.ijmedinf.2021.104643.

Hornback, A., Shi, W., Giuste, F. O., Zhu, Y., Carpenter, A. M., Hilton, C., Bijanki, V. N., Stahl, H., Gottesman, G. S., Purnell, C. et al. (2022). Development of a generalizable multi-site and multi-modality clinical data cloud infrastructure for pediatric patient care. *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–10.

Hossain, M. S., Muhammad, G. & Guizani, N. (2020). Explainable AI and Mass Surveillance System-Based Healthcare Framework to Combat COVID-19 Like Pandemics. *IEEE Network*, 34(4), 126–132. doi: 10.1109/MNET.011.2000051.

Hou, Z., Sun, F., Chen, Y. K., Xie, Y. & Kung, S. Y. (2022). MILAN: Masked Image Pretraining on Language Assisted Representation. *arXiv preprint arXiv:2208.06049*, abs/2208.06049(1), 1–12.

Hu, M., Qian, F., Guo, D., Wang, X., He, L. & Ren, F. (2021a). ETA-rPPGNet: Effective Time-Domain Attention Network for Remote Heart Rate Measurement. *IEEE Transactions on Instrumentation and Measurement*, 70, 1–12. doi: 10.1109/TIM.2021.3051234.

Hu, M., Qian, F., Wang, X., He, L., Guo, D. & Ren, F. (2021b). Robust Heart Rate Estimation With Spatial–Temporal Attention Network From Facial Videos. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2), 639–647. doi: 10.1109/TCDS.2021.3056045.

Hu, M., Guo, D., Wang, X., Ge, P. & Chu, Q. (2019). A Novel Spatial-Temporal Convolutional Neural Network for Remote Photoplethysmography. *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–6. doi: 10.1109/CISP-BMEI48845.2019.8966034.

Hu, M., Qian, F., Guo, D., Wang, X., He, L. & Ren, F. (2021c). ETA-rPPGNet: Effective Time-Domain Attention Network for Remote Heart Rate Measurement. *IEEE Transactions on Instrumentation and Measurement*, 70, 1–12.

Huang, B., Lin, C. L., Chen, W., Juang, C. F. & Wu, X. (2021a). A Novel One-Stage Framework for Visual Pulse Rate Estimation Using Deep Neural Networks. *Biomedical Signal Processing and Control*, 66, 102387. doi: 10.1016/j.bspc.2020.102387.

Huang, B., Hu, S., Liu, Z., Lin, C. L., Su, J., Zhao, C. & Wang, W. (2023). Challenges and Prospects of Visual Contactless Physiological Monitoring in Clinical Study. *NPJ Digital Medicine*, 6(1), 231. doi: 10.1038/s41746-023-00987-5.

Huang, B., Chen, W., Lin, C.-T., Juang, C.-F., Xing, Y., Wang, Y. & Wang, J. (2021b). A Neonatal Dataset and Benchmark for Non-Contact Neonatal Heart Rate Monitoring Based on Spatio-Temporal Neural Networks. *Engineering Applications of Artificial Intelligence*, 106(1), 104447.

Huang, T., Pei, X., You, S., Wang, F., Qian, C. & Xu, C. (2024). Localmamba: Visual state space model with windowed selective scan. *European Conference on Computer Vision*, pp. 12–22.

Hulyalkar, M., Gleich, S. J., Kashyap, R., Barwise, A., Kaur, H., Dong, Y. & Tripathi, S. (2017). Design and α-Testing of an Electronic Rounding Tool (CERTAINp) to Improve Process of Care in Pediatric Intensive Care Unit. *Journal of Clinical Monitoring and Computing*, 31(6), 1313–1320. doi: 10.1007/s10877-016-9940-1.

Huo, J., Wung, S., Roveda, J. & Li, A. (2023). Reducing False Alarms in Intensive Care Units: A Scoping Review. *Exploratory Research and Hypothesis in Medicine*, 8(1), 57–64. doi: 10.14218/ERHM.2022.00072.

Jang, E., Gu, S. & Poole, B. (2016). Categorical reparameterization with gumbel-softmax. Retrieved from: arXivpreprintarXiv:1611.01144.

Kang, J., Yang, S. & Zhang, W. (2024). TransPPG: Two-Stream Transformer for Remote Heart Rate Estimate. *CCF Transactions on Pervasive Computing and Interaction*, 6(3), 271–280. doi: 10.1007/s42486-024-00147-y.

Kaur, M., Marshall, A. P., Eastwood-Sutherland, C., Salmon, B. P., Dargaville, P. A. & Gale, T. J. (2017). Automatic torso detection in images of preterm infants. *Journal of medical systems*, 41(9), 134.

Khanam, F.-T.-Z., Perera, A. G., Al-Naji, A., Gibson, K. & Chahl, J. (2021). Non-contact automatic vital signs monitoring of infants in a neonatal intensive care unit based on neural networks. *Journal of Imaging*, 7(8), 122.

Khemani, R. G., Smith, L. S., Zimmerman, J. J., Erickson, S. & Group, P. A. L. I. C. C. (2015). Pediatric Acute Respiratory Distress Syndrome: Definition, Incidence, and Epidemiology: Proceedings from the Pediatric Acute Lung Injury Consensus Conference. *Pediatric Critical Care Medicine*, 16(5_suppl), S23–S40.

Knop, M., Weber, S., Mueller, M. & Niehaves, B. (2022). Human factors and technological characteristics influencing the interaction of medical professionals with artificial intelligence–enabled clinical decision support systems: literature review. *JMIR Human Factors*, 9(1), e28639.

Krbec, B. A., Zhang, X., Chityat, I., Brady-Mine, A., Linton, E., Copeland, D., Anthony, B. W., Edelman, E. R. & Davis, J. M. (2024). Emerging innovations in neonatal monitoring: a comprehensive review of progress and potential for non-contact technologies. *Frontiers in Pediatrics*, 12, 1442753. doi: 10.3389/fped.2024.1442753.

Kumar, M., Veeraraghavan, A. & Sabharwal, A. (2015). DistancePPG: Robust non-contact vital signs monitoring using a camera. *Biomedical Optics Express*, 6(5), 1565–1588. doi: 10.1364/BOE.6.001565.

Kyriacou, P. A. & Allen, J. (Eds.). (2021). *Photoplethysmography: Technology, Signal Analysis and Applications*. London: Academic Press.

Kyrollos, D. G., Tanner, J. B., Greenwood, K., Harrold, J. & Green, J. R. (2021). Noncontact Neonatal Respiration Rate Estimation Using Machine Vision. *Proceedings of the 2021 IEEE Sensors Applications Symposium (SAS)*, 1(1), 1–6.

Lee, E., Chen, E. & Lee, C. Y. (2020a). Meta-rPPG: Remote Heart Rate Estimation Using a Transductive Meta-learner. In Vedaldi, A., Bischof, H., Brox, T. & Frahm, J. M. (Eds.), *Computer Vision – ECCV 2020* (vol. 12372, pp. 392–409). Cham: Springer. doi: 10.1007/978-3-030-58583-9_24.

Lee, E., Chen, E. & Lee, C.-Y. (2020b). Meta-rPPG: Remote Heart Rate Estimation Using a Transductive Meta-learner. *Computer Vision – ECCV 2020*, 392–409.

Lee, S., Lee, M. & Sim, J. Y. (2023a). DSE-NN: Deeply Supervised Efficient Neural Network for Real-Time Remote Photoplethysmography. *Bioengineering*, 10(12), 1428. doi: 10.3390/bioengineering10121428.

Lee, S., Lee, M. & Sim, J. Y. (2023b). DSE-NN: Deeply Supervised Efficient Neural Network for Real-Time Remote Photoplethysmography. *Bioengineering*, 10(12), 1428.

Li, F., Zhang, L., Liu, Z., Lei, J. & Li, Z. (2023a). Multi-frequency representation enhancement with privilege information for video super-resolution. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12814–12825.

Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L. & Qiao, Y. (2024). VideoMamba: State Space Model for Efficient Video Understanding. In *European Conference on Computer Vision (ECCV)* (pp. 237–255). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-72636-2_14.

Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L. & Qiao, Y. (2023b). Unmasked teacher: Towards training-efficient video foundation models. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19948–19960.

Li, Y., Mao, H., Girshick, R. & He, K. (2022). Exploring plain vision transformer backbones for object detection. *European conference on computer vision*, pp. 280–296.

Li, Z. & Yin, L. (2023). Contactless Pulse Estimation Leveraging Pseudo Labels and Self-Supervision. In 2023 IEEE. *CVF International Conference on Computer Vision (ICCV)*, pp. 20531–20540.

Liang, Y., Chen, Z., Ward, R. & Elgendi, M. (2018). Photoplethysmography and Deep Learning: Enhancing Hypertension Risk Stratification. *Biosensors*, 8(4), 101. doi: 10.3390/bios8040101.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. *Proceedings of the European Conference on Computer Vision (ECCV)*, 8693(1), 740–755.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B. & Sánchez, C. I. (2017). A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42(1), 60–88.

Liu, J., Yang, H., Zhou, H.-Y., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S. et al. (2024a). Swin-umamba: Mamba-based unet with imagenet-based pretraining. *International conference on medical image computing and computer-assisted intervention*, pp. 615–625.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J. & Han, J. (2019). On the Variance of the Adaptive Learning Rate and Beyond. *arXiv preprint arXiv:1908.03265*, abs/1908.03265(1), 1–15.

Liu, M., Tang, J., Chen, Y., Li, H., Qi, J., Li, S. & Chen, H. (2025). Spiking-PhysFormer: Camera-Based Remote Photoplethysmography With Parallel Spike-Driven Transformer. *Neural Networks*, 185, 107128. doi: 10.1016/j.neunet.2024.107128.

Liu, S. Q. & Yuen, P. C. (2020). A General Remote Photoplethysmography Estimator With Spatiotemporal Convolutional Network. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 481–488. doi: 10.1109/FG47880.2020.00075.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. & Berg, A. C. (2016). SSD: Single Shot Multibox Detector. *Proceedings of the European Conference on Computer Vision (ECCV)*, 9905(1), 21–37.

Liu, X., Fromm, J., Patel, S. & McDuff, D. (2020). Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33, 19400–19411.

Liu, X., Zhang, Y., Yu, Z., Lu, H., Yue, H. & Yang, J. (2024b). rPPG-MAE: Self-Supervised Pretraining With Masked Autoencoders for Remote Physiological Measurements. *IEEE Transactions on Multimedia*, 26, 7278–7293. doi: 10.1109/TMM.2024.3362935.

Liu, X., Hill, B. L., Jiang, Z., Patel, S. & McDuff, D. (2022a). *EfficientPhys: Enabling Simple, Fast and Accurate Camera-Based Vitals Measurement*. arXiv preprint arXiv:2110.04447. Retrieved from: https://arxiv.org/abs/2110.04447.

Liu, X., Wei, W., Kuang, H. & Ma, X. (2022b). Heart Rate Measurement Based on 3D Central Difference Convolution with Attention Mechanism. *Sensors*, 22(2), 688.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J. & Liu, Y. (2024c). Vmamba: Visual state space model. *Advances in neural information processing systems*, 37, 103031–103063.

Lu, H., Salah, A. A. & Poppe, R. (2024). Videomambapro: A leap forward for mamba in video understanding. *arXiv e-prints*, arXiv–2406.

Luo, C., Xie, Y. & Yu, Z. (2024). PhysMamba: Efficient Remote Physiological Measurement With SlowFast Temporal Difference Mamba. In *Chinese Conference on Biometric Recognition* (pp. 248–259). Singapore: Springer Nature Singapore. doi: 10.1007/978-981-97-1018-2_21.

L'Her, E., Nazir, S., Pateau, V. & Visvikis, D. (2022). Accuracy of noncontact surface imaging for tidal volume and respiratory rate measurements in the ICU. *Journal of Clinical Monitoring and Computing*, 36(3), 775–783.

Massaroni, C., Lo Presti, D., Formica, D., Silvestri, S. & Schena, E. (2019). Non-contact monitoring of breathing pattern and respiratory rate via RGB signal measurement. *Sensors*, 19(12), 2758.

McNichol, L., Lund, C., Rosen, T. & Gray, M. (2013). Medical Adhesives and Patient Safety: State of the Science: Consensus Statements for the Assessment, Prevention, and Treatment of Adhesive-Related Skin Injuries. *Journal of the Dermatology Nurses' Association*, 5(6), 323–338. doi: 10.1097/JDN.0000000000000007.

Mehta, H., Gupta, A., Cutkosky, A. & Neyshabur, B. (2022). Long Range Language Modeling via Gated State Spaces. *arXiv preprint arXiv:2206.13947*, abs/2206.13947(1), 1–10.

Moraes, J. L., Rocha, M. X., Vasconcelos, G. G., Vasconcelos Filho, J. E., De Albuquerque, V. H. C. & Alexandria, A. R. (2018). Advances in photoplethysmography signal analysis for biomedical applications. *Sensors*, 18(6), 1894. doi: 10.3390/s18061894.

Muralidharan, V., Burgart, A., Daneshjou, R. & Rose, S. (2023). Recommendations for the Use of Pediatric Data in Artificial Intelligence and Machine Learning ACCEPT-AI. *NPJ Digital Medicine*, 6(1), 166. doi: 10.1038/s41746-023-00954-0.

Nagar, S., Hasegawa-Johnson, M., Beiser, D. G. & Ahuja, N. (2024). *R2I-rPPG: A Robust Region of Interest Selection Method for Remote Photoplethysmography to Extract Heart Rate*. arXiv preprint arXiv:2410.15851. Retrieved from: https://arxiv.org/abs/2410.15851.

Nagy, Á., Földesy, P., Jánoki, I., Terbe, D., Siket, M., Szabó, M., Varga, J. & Zarándy, Á. (2021). Continuous camera-based premature-infant monitoring algorithms for NICU. *Applied Sciences*, 11(16), 7215.

Nguyen, N., Nguyen, L., Li, H., López, M. B. & Casado, C. Á. (2024). Evaluation of Video-Based rPPG in Challenging Environments: Artifact Mitigation and Network Resilience. *Computers in Biology and Medicine*, 179, 108873. doi: 10.1016/j.compbiomed.2024.108873.

Nisevic, M., Milojevic, D. & Spajic, D. (2025). Synthetic Data in Medicine: Legal and Ethical Considerations for Patient Profiling. *Computational and Structural Biotechnology Journal*, 28, 190–198. doi: 10.1016/j.csbj.2025.05.026.

Niu, X., Han, H., Shan, S. & Chen, X. (2018). VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. *Asian conference on computer vision*, pp. 562–576.

Niu, X., Shan, S., Han, H. & Chen, X. (2020a). RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation. *IEEE Transactions on Image Processing*, 29, 2409–2423. doi: 10.1109/TIP.2019.2947204.

Niu, X., Shan, S., Han, H. & Chen, X. (2020b). RhythmNet: End-to-end Heart Rate Estimation from Face via Spatial-temporal Representation. *IEEE Transactions on Image Processing*, 29, 2409–2423.

Nowara, E. M., McDuff, D. & Veeraraghavan, A. (2020). A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 284–285.

Park, J., Seok, H. S., Kim, S. S. & Shin, H. (2022). Photoplethysmogram analysis and applications: an integrative review. *Frontiers in Physiology*, 12, 808451. doi: 10.3389/fphys.2021.808451.

Parshuram, C. S. & et al. (2011). Multicentre Validation of the Bedside Paediatric Early Warning System Score: A Severity of Illness Score to Detect Evolving Critical Illness in Hospitalised Children. *Critical Care*, 15(4), R184. doi: 10.1186/cc10337.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Chintala, S. et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32(1), 8024–8035.

Patel, V. L. & Buchman, T. G. (2016). Cognitive Overload in the ICU [Online resource]. Retrieved from: https://psnet.ahrq.gov/perspective/cognitive-overload-icu.

Pei, X., Huang, T. & Xu, C. (2025). Efficientvmamba: Atrous selective scan for light weight visual mamba. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6), 6443–6451.

Perepelkina, O., Artemyev, M., Churikova, M. & Grinenko, M. (2020). HeartTrack: Convolutional neural network for remote video-based heart rate monitoring. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1163–1171.

Pióro, M., Ciebiera, K., Król, K., Ludziejewski, J., Krutul, M., Krajewski, J. & Jaszczur, S. (2024). MoE-Mamba: Efficient Selective State Space Models with Mixture of Experts. *arXiv preprint arXiv:2401.04081*, abs/2401.04081(1), 1–12.

Poh, M. Z., McDuff, D. J. & Picard, R. W. (2010a). Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10), 10762–10774. doi: 10.1364/OE.18.010762.

Poh, M.-Z., McDuff, D. J. & Picard, R. W. (2010b). Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10), 10762–10774.

Popoff, B., Cabon, S., Cuggia, M., Bouzillé, G. & Clavier, T. (2025). Expectations of Intensive Care Physicians Regarding an AI-Based Decision Support System for Weaning From Continuous Renal Replacement Therapy: Predevelopment Survey Study. *JMIR Medical Informatics*, 13(1), e63709. doi: 10.2196/63709.

Qi, D., Tan, W., Yao, Q. & Liu, J. (2022). YOLO5Face: Why Reinventing a Face Detector. *Proceedings of the European Conference on Computer Vision (ECCV)*, 13671(1), 228–244.

Qiu, Y., Liu, Y., Arteaga-Falconi, J., Dong, H. & El Saddik, A. (2019a). EVM-CNN: Real-Time Contactless Heart Rate Estimation From Facial Video. *IEEE Transactions on Multimedia*, 21(7), 1778–1787. doi: 10.1109/TMM.2018.2883866.

Qiu, Y., Liu, Y., Arteaga-Falconi, J., Dong, H. & Saddik, A. E. (2019b). EVM-CNN: Real-Time Contactless Heart Rate Estimation From Facial Video. *IEEE Transactions on Multimedia*, 21(7), 1778–1787.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, pp. 8748–8763.

Redmon, J. & Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1(1), 7263–7271.

Redmon, J. & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767*, abs/1804.02767(1), 1–6.

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1(1), 779–788.

Rehouma, H., Noumeir, R., Bouachir, W., Jouvet, P. & Essouri, S. (2018). 3D imaging system for respiratory monitoring in pediatric intensive care environment. *Computerized Medical Imaging and Graphics*, 70, 17–28.

Ren, S., He, K., Girshick, R. & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, 28(1), 91–99.

Revanur, A., Dasari, A., Tucker, C. S. & Jeni, L. A. (2022). Instantaneous Physiological Estimation Using Video Transformers. In *Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence* (pp. 307–319). Cham: Springer International Publishing. doi: 10.1007/978-3-031-20862-2_20.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K. et al. (2020). The future of digital health with federated learning. *NPJ digital medicine*, 3(1), 119.

Rohmetra, H., Raghunath, N., Narang, P., Chamola, V., Guizani, M. & Lakkaniga, N. R. (2023). AI-Enabled Remote Monitoring of Vital Signs for COVID-19: Methods, Prospects and Challenges. *Computing*, 105(4), 783–809. doi: 10.1007/s00607-022-01175-1.

Rong, Y. & Bliss, D. W. (2023). Insights on using time-of-flight camera for recovering cardiac pulse from chest motion in depth videos. *IEEE Transactions on Biomedical Engineering*, 71(3), 772–779.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241.

Ruan, J., Li, J. & Xiang, S. (2024). VM-UNet: Vision Mamba UNet for Medical Image Segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 1(1), 1–10.

Ruhrberg Estévez, S., Grafton, A., Thomson, L., Warnecke, J., Beardsall, K. & Lasenby, J. (2025). Continuous non-contact vital sign monitoring of neonates in intensive care units using RGB-D cameras. *Scientific Reports*, 15(1), 16863. doi: 10.1038/s41598-025-00539-9.

Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J. & Churpek, M. M. (2018). Comparison of Variable Selection Methods for Clinical Predictive Modeling. *International Journal of Medical Informatics*, 116, 10–17. doi: 10.1016/j.ijmedinf.2018.05.006.

Savic, M. & Zhao, G. (2024a). Physu-net: Long temporal context transformer for rppg with self-supervised pre-training. *International Conference on Pattern Recognition*, pp. 228–243.

Savic, M. & Zhao, G. (2024b). Rs-rppg: Robust self-supervised learning for rppg. *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–10.

Senechal, E., Radeschi, D., Tao, L., Lv, S., Jeanne, E., Kearney, R., Shalish, W. & Sant Anna, G. (2023). The use of wireless sensors in the neonatal intensive care unit: a study protocol. *PeerJ*, 11, e15578. doi: 10.7717/peerj.15578.

Shen, W., Zhou, M., Yang, F., Yang, C. & Tian, J. (2015). Multi-scale convolutional neural networks for lung nodule classification. *International conference on information processing in medical imaging*, pp. 588–599.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k. & Woo, W.-c. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems*, 28(1), 802–810.

Singh, P., Nampalle, K. B., Narayan, U. V. & Raman, B. (2023). *See Through the Fog: Curriculum Learning With Progressive Occlusion in Medical Imaging*. arXiv preprint arXiv:2306.15574. Retrieved from: https://arxiv.org/abs/2306.15574.

Smith, J. T., Warrington, A. & Linderman, S. W. (2022). Simplified State Space Layers for Sequence Modeling. *arXiv preprint arXiv:2208.04933*, abs/2208.04933(1), 1–10.

Snyder, K. B., Stewart, R. A. & Hunter, C. J. (2025). Ethical Considerations for the Application of Artificial Intelligence in Pediatric Surgery. *AI and Ethics*, 5(2), 1885–1892. doi: 10.1007/s43681-024-00356-5.

Song, R., Chen, H., Cheng, J., Li, C., Liu, Y. & Chen, X. (2021). PulseGAN: Learning to Generate Realistic Pulse Waveforms in Remote Photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 25(5), 1373–1384. doi: 10.1109/JBHI.2020.3028602.

Speth, J., Vance, N., Flynn, P. & Czajka, A. (2023). Non-contrastive unsupervised learning of physiological signals from video. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14464–14474.

Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P. & Vaswani, A. (2021). Bottleneck Transformers for Visual Recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16514–16524.

Sun, W., Zhang, X., Lu, H., Chen, Y., Ge, Y., Huang, X. & Chen, Y. (2023). Resolve Domain Conflicts for Generalizable Remote Physiological Measurement. *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8214–8224. doi: 10.1145/3581783.3612265.

Sun, Z. & Li, X. (2022). Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. *European Conference on Computer Vision*, pp. 492–510.

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

Svoboda, L., Sperrhake, J., Nisser, M., Taphorn, L. & Proquitté, H. (2024). Contactless Assessment of Heart Rate in Neonates Within a Clinical Environment Using Imaging Photoplethysmography. *Frontiers in Pediatrics*, 12, 1383120. doi: 10.3389/fped.2024.1383120.

Takano, C. & Ohta, Y. (2007). Heart rate measurement based on a time-lapse image. *Medical Engineering & Physics*, 29(8), 853–857. doi: 10.1016/j.medengphy.2006.09.006.

Talukdar, D., de Deus, L. F. & Sehgal, N. (2023). Evaluation of Remote Monitoring Technology across different skin tone participants. *medRxiv*, 2023–04.

Tong, Z., Song, Y., Wang, J. & Wang, L. (2022). VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. *Advances in Neural Information Processing Systems*, 35(1), 10078–10093.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1(1), 4489–4497.

Tsou, Y.-Y., Lee, Y.-A., Hsu, C.-T. & Chang, S.-H. (2020). Siamese-rPPG network: remote photoplethysmography signal estimation from face videos. *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2066–2073.

U.S. Food and Drug Administration. (2019). *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)*. Retrieved from: https://www.fda.gov/media/122535/download.

U.S. Food and Drug Administration. (2021). *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*. Retrieved from: https://www.fda.gov/media/145022/download.

U.S. Food and Drug Administration. (2025). *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions*. Retrieved from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence.

van Rossum, M. C., Vlaskamp, L. B., Posthuma, L. M., Visscher, M. J., Breteler, M. J., Hermens, H. J. & Preckel, B. (2022). Adaptive Threshold-Based Alarm Strategies for Continuous Vital Signs Monitoring. *Journal of Clinical Monitoring and Computing*, 36(2), 407–417. doi: 10.1007/s10877-021-00664-2.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems*, 30(1), 5998–6008.

Verkruysse, W., Svaasand, L. O. & Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26), 21434–21445. doi: 10.1364/OE.16.021434.

Villarroel, M., Chaichulee, S., Jorge, J., Davis, S., Green, G., Arteta, C. & Tarassenko, L. (2019). Non-Contact Physiological Monitoring of Preterm Infants in the Neonatal Intensive Care Unit. *NPJ Digital Medicine*, 2(1), 128. doi: 10.1038/s41746-019-0201-5.

Špetlík, R., Franc, V. & Matas, J. (2018, September). Visual heart rate estimation with convolutional neural network. *Proceedings of the British Machine Vision Conference*, pp. 3–6.

Wang, A., Islam, M., Xu, M. & Ren, H. (2024a). Curriculum-Based Augmented Fourier Domain Adaptation for Robust Medical Image Segmentation. *IEEE Transactions on Automation Science and Engineering*, 21(3), 4340–4352. doi: 10.1109/TASE.2023.3295600.

Wang, C., Tsepa, O., Ma, J. & Wang, B. (2024b). Graph-Mamba: Towards Long-Range Graph Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2402.00789*, abs/2402.00789(1), 1–10.

Wang, G., Li, B., Zhang, T. & Zhang, S. (2022a). A Network Combining a Transformer and a Convolutional Neural Network for Remote Sensing Image Change Detection. *Remote Sensing*, 14(9), 2228.

Wang, H., Ahn, E. & Kim, J. (2022b). Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2), 2431–2439.

Wang, J., Shan, C., Liu, Z., Zhou, S. & Shu, M. (2025). Physiological Information Preserving Video Compression for rPPG. *IEEE Journal of Biomedical and Health Informatics*, 1(1), 1–10.

Wang, R.-X., Sun, H.-M., Hao, R.-R., Pan, A. & Jia, R.-S. (2023). TransPhys: Transformer-based unsupervised contrastive learning for remote heart rate measurement. *Biomedical Signal Processing and Control*, 86, 105058.

Wang, W., den Brinker, A. C., Stuijk, S. & de Haan, G. (2017a). Algorithmic Principles of Remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7), 1479–1491. doi: 10.1109/TBME.2016.2609282.

Wang, W., den Brinker, A. C., Stuijk, S. & de Haan, G. (2017b). Algorithmic Principles of Remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7), 1479–1491.

Wang, X., Girshick, R., Gupta, A. & He, K. (2018). Non-Local Neural Networks. *Conference on Computer Vision and Pattern Recognition*, 7794–7803.

Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. *European Conference on Computer Vision*, 3–19.

Wu, Z., Xie, Y., Zhao, B., He, J., Luo, F., Deng, N. & Yu, Z. (2025). *CardiacMamba: A Multimodal RGB-RF Fusion Framework With State Space Models for Remote Physiological Measurement*. arXiv preprint arXiv:2502.13624. Retrieved from: https://arxiv.org/abs/2502.13624.

Xiao, H., Li, Z., Xia, Z., Liu, T., Zhou, F. & Avolio, A. (2024). SimFuPulse: A self-similarity supervised model for remote photoplethysmography extraction from facial videos. *Biomedical Signal Processing and Control*, 98, 106736.

Xie, Y., Yu, Z., Wu, B., Xie, W. & Shen, L. (2024). *SFDA-rPPG: Source-Free Domain Adaptive Remote Physiological Measurement With Spatio-Temporal Consistency*. arXiv preprint arXiv:2409.12040. Retrieved from: https://arxiv.org/abs/2409.12040.

Xing, Z., Ye, T., Yang, Y., Liu, G. & Zhu, L. (2024). Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *International conference on medical image computing and computer-assisted intervention*, pp. 578–588.

Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D. & Zhuang, Y. (2019). Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1(1), 10334–10343.

Yang, J., Soltan, A. A. & Clifton, D. A. (2022a). Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. *NPJ digital medicine*, 5(1), 69.

Yang, J., Soltan, A. A. & Clifton, D. A. (2022b). Machine Learning Generalizability Across Healthcare Settings: Insights from Multi-Site COVID-19 Screening. *NPJ Digital Medicine*, 5(1), 69.

Yu, Z., Li, X. & Zhao, G. (2019a). *Remote Photoplethysmograph Signal Measurement From Facial Videos Using Spatio-Temporal Networks*. arXiv preprint arXiv:1905.02419. Retrieved from: https://arxiv.org/abs/1905.02419.

Yu, Z., Li, X., Niu, X., Shi, J. & Zhao, G. (2020). AutoHR: A Strong End-to-End Baseline for Remote Heart Rate Measurement With Neural Searching. *IEEE Signal Processing Letters*, 27, 1245–1249. doi: 10.1109/LSP.2020.3003299.

Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P. H. & Zhao, G. (2022). PhysFormer: Facial Video-Based Physiological Measurement With Temporal Difference Transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4186–4196.

Yu, Z., Shen, Y., Shi, J., Zhao, H., Cui, Y., Zhang, J. & Zhao, G. (2023). PhysFormer++: Facial Video-Based Physiological Measurement With SlowFast Temporal Difference Transformer. *International Journal of Computer Vision*, 131(6), 1307–1330. doi: 10.1007/s11263-023-01798-6.

Yu, Z., Li, X. & Zhao, G. (2019b). Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. *arXiv preprint arXiv:1905.02419*, abs/1905.02419(1), 1–10.

Yu, Z., Peng, W., Li, X., Hong, X. & Zhao, G. (2019c). Remote Heart Rate Measurement From Highly Compressed Facial Videos: An End-to-End Deep Learning Solution With Video Enhancement. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 151–160. doi: 10.1109/ICCV.2019.00024.

Yu, Z., Peng, W., Li, X., Hong, X. & Zhao, G. (2019d). Remote Heart Rate Measurement From Highly Compressed Facial Videos: An End-to-End Deep Learning Solution With Video Enhancement. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 151–160.

Yu, Z., Zhou, B., Wan, J., Wang, P., Chen, H., Liu, X., Li, S. Z. & Zhao, G. (2021). Searching Multi-Rate and Multi-Modal Temporal Enhanced Networks for Gesture Recognition. *IEEE Transactions on Image Processing*, 30, 5626–5640.

Yuan, X., Zhang, S., Zhao, C., He, X., Ouyang, B. & Yang, S. (2022). Pain Intensity Recognition from Masked Facial Expressions Using Swin-Transformer. *Proceedings of the 2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 1(1), 723–728.

Yue, Z., Shi, M. & Ding, S. (2023). Facial video-based remote physiological measurement via self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11), 13844–13859.

Zhang, K., Zhang, Z., Li, Z. & Qiao, Y. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10), 1499–1503.

Zhang, N., Sun, H.-M., Ma, J.-R. & Jia, R.-S. (2024). A self-supervised learning network for remote heart rate measurement. *Measurement*, 228, 114379.

Zhang, X., Yang, C., Yin, R. & Meng, L. (2023a). An End-to-End Heart Rate Estimation Scheme Using Divided Space-Time Attention. *Neural Processing Letters*, 55(3), 2661–2685. doi: 10.1007/s11063-022-11164-4.

Zhang, X., Yang, C., Yin, R. & Meng, L. (2023b). An End-to-End Heart Rate Estimation Scheme Using Divided Space-Time Attention. *Neural Processing Letters*, 55(3), 2661–2685.

Zhang, Y., Lu, H., Liu, X., Chen, Y. & Wu, K. (2025). Advancing Generalizable Remote Physiological Measurement Through the Integration of Explicit and Implicit Prior Knowledge. *IEEE Transactions on Image Processing*, 34, 3764–3778. doi: 10.1109/TIP.2025.3576490.

Zhang, Z., Cheng, D., Zhu, X., Lin, S. & Dai, J. (2018). Integrated Object Detection and Tracking with Tracklet-Conditioned Detection. *arXiv preprint arXiv:1811.11167*, abs/1811.11167(1), 1–10.

Zhao, H., Zhang, M., Zhao, W., Ding, P., Huang, S. & Wang, D. (2025). Cobra: Extending mamba to multi-modal large language model for efficient inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10), 10421–10429.

Zhao, Y., Zou, B., Yang, F., Lu, L., Belkacem, A. N. & Chen, C. (2021). Video-Based Physiological Measurement Using 3D Central Difference Convolution Attention Network. *2021 IEEE International Joint Conference on Biometrics (IJCB)*, 1–6.

Zheng, F., Chen, X., Liu, W., Li, H., Lei, Y., He, J. & Zhou, S. (2024). Smaformer: Synergistic Multi-Attention Transformer for Medical Image Segmentation. *Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1(1), 4048–4053.

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W. & Wang, X. (2024). Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *arXiv preprint arXiv:2401.09417*, abs/2401.09417(1), 1–12.

Zhu, M. & Gupta, S. (2017). To prune, or not to prune: exploring the efficacy of pruning for model compression.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X. & Dai, J. (2020). Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv preprint arXiv:2010.04159*, abs/2010.04159(1), 1–10.

Zou, B., Guo, Z., Hu, X. & Ma, H. (2024). Rhythmmamba: Fast remote physiological measurement with arbitrary length videos.

Zou, B., Guo, Z., Chen, J., Zhuo, J., Huang, W. & Ma, H. (2025). RhythmFormer: Extracting Patterned rPPG Signals Based on Periodic Sparse Attention. *Pattern Recognition*, 164, 111511. doi: 10.1016/j.patcog.2025.111511.

Álvarez Casado, C., Nguyen, L., Silvén, O. & Bordallo López, M. (2023a). Assessing the Feasibility of Remote Photoplethysmography through Videocalls: A Study of Network and Computing Constraints. *Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, 13919(1), 586–598.

Álvarez Casado, C., Padilla-López, J. R., Latorre-Crespo, E., Pérez-Borràs, A. & Casas, J. R. (2023b). Face2PPG: An unsupervised pipeline for blood volume pulse extraction from faces. *Sensors*, 23(5), 2466. doi: 10.3390/s23052466.

Špetlík, R., Franc, V. & Matas, J. (2018). Visual Heart Rate Estimation with Convolutional Neural Network. *Proceedings of the British Machine Vision Conference (BMVC)*, 1(1), 3–6.