CliffPhys: Camera-based Respiratory Measurement using Clifford Neural Networks

Omar Ghezzi[1], Giuseppe Boccignone[1], Giuliano Grossi[1], Raffaella Lanzarotti[1], and Alessandro D'Amelio[1]

PHuSe Lab - Università degli Studi di Milano
omar.ghezzi@studenti.unimi.it
{giuseppe.boccignone,giuliano.grossi,
raffaella.lanzarotti,alessandro.damelio}@unimi.it

Abstract. This paper presents CliffPhys, a family of models that leverage hypercomplex neural architectures for camera-based respiratory measurement. The proposed approach extracts respiratory motion from standard RGB cameras, relying on optical flow and monocular depth estimation to obtain a 2D vector field and a scalar field, respectively. We show how the adoption of Clifford Neural Layers to model the geometric relationships within the recovered input fields allows respiratory information to be effectively estimated. Experimental results on three publicly available datasets demonstrate CliffPhys' superior performance compared to both baselines and recent neural approaches, achieving state-of-the-art results in the prediction of respiratory rates. Source code available at: https://github.com/phuselab/CliffPhys.

Keywords: Contactless Respiration Monitoring · Clifford Neural Layers · Remote Physiological Measurement · Vital Signs Monitoring

1 Introduction

Camera-based physiological measurement refers to a relatively recent field of research borrowing expertise from computer vision, machine learning, signal processing and biomedical engineering to deliver techniques allowing a non-invasive estimation of physiological signals through common cameras. In general, obtaining vital sign measurements necessitates invasive equipment positioned appropriately on the subject's body. This requirement often deters the widespread adoption of such data due to several factors, ranging from physical discomfort to the practical impossibility of positioning the required sensors [5].

In recent decades, the development of contactless solutions has facilitated the measurement of various physiological signals, including blood volume pulse (BVP), blood pressure, electro-dermal activity (EDA), blood oxygenation levels (SpO_2) and respiratory signals/rates (for a recent review see [28]). Albeit latest trends have explored self-supervised or unsupervised learning schemes [45, 54], most recent literature claims supervised neural approaches as the most reliable methods for camera-based vital sign measurement [24]. However, the difficulty

in collecting large-scale datasets that provide physiological ground truths, combined with the necessity of light solutions eventually able to run on-device [23], has pushed the community to devise efficient architectures with extensive adoption of inductive biases [7, 22]. In such regard, respiratory measurement makes no exception. In light of this, the present work proposes a novel solution for camera-based respiratory measurement that builds upon most recent hypercomplex neural approaches capable of exploiting geometric inductive biases favoring the extraction of respiratory motion from standard RGB cameras. The results show how the proposed family of methods is able to extract subtle respiratory motion from the estimated depth representation and the apparent motion, both obtained from the original RGB frames. Indeed, the adoption of hypercomplex neural networks, specifically Clifford Neural layers, enables a principled mathematical treatment of the obtained representation. In summary, the contributions of this work are: **1)** we propose the adoption of optical flow and monocular depth estimation as a video pre-processing step providing strong inductive bias for respiratory motion extraction. The obtained representation is mathematically treated as the combination of two different kinds of fields: a scalar field (depth) and a 2D vector field (apparent motion). **2)** A family of novel hypercomplex neural architectures (`CliffPhys`) based on the recently proposed Clifford Neural Layers is presented. Such architectures allow us to explicitly model the geometrical relationships between the above-mentioned input fields.

To validate our approach, we conducted experiments on three datasets, namely SCAMPS [29] (used in the pre-training step), COHFACE [14] and BP4D+ [58]. Our methods outperform all the adopted baselines as well as recent neural approaches in terms of accuracy in the prediction of respiratory rates, thus setting SOTA results on all the adopted benchmark datasets.

2 Related Work

2.1 Camera-based Respiratory Measurement

Current RGB camera-based respiratory rate estimation methods fall into two categories: rPPG-based and motion-based. The first exploits the link between cardiac and respiratory activity, targeting the extraction of respiratory-induced variations (RIVs) in the estimated cardiac signal [2, 11, 26, 37, 49, 50]. Conversely, motion-based approaches track respiratory activity by analyzing the tiny, periodic variations in pixel intensity values caused by lungs movement. Techniques like optical flow (OF) are commonly used for this purpose [17, 21, 27, 38, 52]. More recently, various neural architectures have been employed to predict respiratory information from videos, such as autoencoders [33], LSTMs [19] or (spatio-temporal) CNNs [4, 16, 46, 51]. Currently, the most popular solutions employ dual-branch architectures, modelling motion and appearance simultaneously [6]. The motion branch analyzes the difference between consecutive face crops, while the appearance branch generates a soft-attention mask (SAM) highlighting skin regions crucial for physiological signal extraction. Along the same lines, [22] introduces a multi-task temporal shift convolutional attention network (`MTTS-CAN`) to

predict cardiovascular and respiratory measurements simultaneously. Similar solutions have been proposed in [20,41,42]. More recently, Narayanswamy et al. [31] introduced `BigSmall`, a model that simultaneously estimates facial expressions, heart rate, and respiratory rate. It implements a high-resolution branch (Big) for spatial features and a low-resolution branch (Small) for temporal dynamics. In a different vein, Yu et al. [55,56] explored transformer architectures for rPPG signal extraction and respiration rate estimation.

2.2 Hypercomplex Neural Networks

Hypercomplex neural networks, leveraging complex, quaternion, and Clifford algebraic structures, have attracted considerable interest in diverse fields, such as computer vision [36], NLP [47], and partial differential equations (PDE) solving [3], demonstrating promising results while reducing model parameter size. Danihelka et al. [9] introduced Associative LSTMs with complex vectors, improving memory and learning for multiple memorization tasks without extra parameters. Chiheb et al. [8] proposed Deep Complex Networks (DCNs) with complex-valued convolutions, achieving competitive performance in computer vision, music, and speech tasks compared to real-valued models. Quaternion-recurrent neural networks (QRNN) have been presented in [35]. QRNNs and LSTMs leverage quaternion algebra for superior speech recognition, reducing parameters compared to traditional models. Similarly, Zhang et al. [57] demonstrated improved performance and flexibility with hypercomplex neural networks compared to real-valued LSTM and transformer counterparts. More recently, Brandstetter et al. [3] introduced Clifford Neural Layers with multivector representations and convolutions for scientific simulations, achieving better generalization. This line of work has been extended with Geometric Clifford Algebra Networks (GCANs) [44] and Clifford Group Equivariant Neural Networks [43].

3 Background and Motivation

3.1 Camera-based Respiratory Motion Model

The respiratory process causes the lungs to mechanically engage, producing changes in intrathoracic volume that translate to movement of the subject's head, chest, and thorax. Camera-based respiratory motion extraction methods aim at estimating the subtle respiratory-related spatio-temporal variations eventually described by the pixel intensity values captured by a camera sensor.

Specifically, let $F_t(i,j)$ represent the light intensity of the video frame at time t and spatial coordinates (i,j); in the vein of [51], the camera signal can be approximated as:

$$F_t(i,j) \approx I_t(i,j)R_t(i,j) = I_t(i,j)M_t(i-\nu^{(i)}t, j-\nu^{(j)}t) \quad , \tag{1}$$

where I represents the (time-varying) illumination strength and R is the amount of reflected light hitting the camera sensor. The latter can be rewritten in terms

of the displacement of objects, $M_t(i-\nu^{(i)}t, j-\nu^{(j)}t)$, inside the scene due to the apparent motion that originated from arbitrary or respiratory movements. The vast majority of motion-based approaches for respiratory extraction aim at disentangling respiratory induced motion from the movement pattern represented by $\nu^{(j)}$, hence assuming that relevant information (respiration) appears, by and large, as rigid vertical motion. However, it should be noted that intrathoracic volume variations appear as vertical as well as medial-lateral and anterior-posterior changes [32]. Interestingly enough, various approaches have explored the adoption of depth cameras to estimate respiratory rates and volumes [18, 34], thus exploiting anterior-posterior chest variations. Clearly, leveraging all three kinds of motion may result in more robust estimates; consequently, we extend the model in Eq. (1) to consider the depth information:

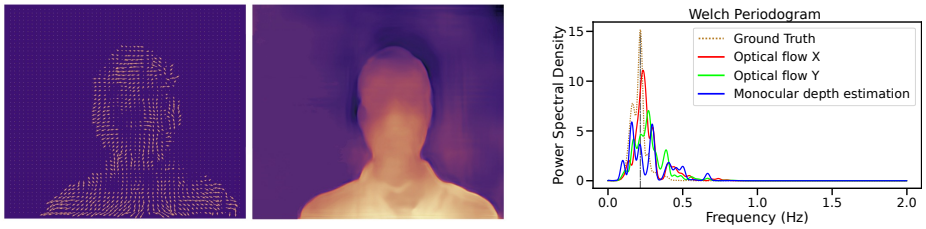
$$F_t(i,j,z) = I_t(i,j)M_t(i-\nu^{(i)}t, j-\nu^{(j)}t, z) \quad . \tag{2}$$

Depth maps (z), can be obtained using RGB-D cameras or estimated monocularly. For instance, Fig. 1a depicts the 2D vector field estimated via optical flow on a video from the COHFACE dataset, along with the corresponding depth map estimated via Monocular Depth Estimation (MDE). Fig. 1b shows the Power Spectral Densities (PSDs) of the bandpass-filtered signals obtained by averaging the motion vectors in the horizontal or vertical direction (across the frames of a video), and the average depth map (across time). The PSD of ground truth subject's respiratory signal is depicted with a dashed line. As can be qualitatively observed, the three signals carry some respiratory information that can be combined to achieve more robust predictions. Taking Fig. 1a as our motivating example and inspired by [3], we observe that the respiratory induced motion patterns can be spatio-temporally described by a vector field (2D apparent motion) and a scalar field (depth map). Such quantities are strongly connected as describing the same physical process.

Conventional deep learning methods treat vector field components similarly to scalar fields, aggregating all fields along the channel dimension. This approach often overlooks geometric relationships among different components, neglecting crucial inductive biases inherent in the data. In our context, optical flow represents a vector field point, while depth information relates as a scalar linked to the two-dimensional respiratory motion space. Clifford Algebra networks offer a natural representation integrating scalar, hypercomplex, and vector components, constructing the learned predictor as the optimal series of geometric transformations within the data space.

3.2 Clifford Algebra and Clifford Neural Layers

Clifford Algebra over $\mathbb{R}$. Clifford Algebra over the real field $\mathbb{R}$ enriches $\mathbb{R}^n$ with a bilinear operator called geometric or Clifford product [10, 25]. Specifically, given two non-negative integers, p and q, where their sum satisfies $p+q=n$, the real Clifford algebra $Cl_{p,q}$ with signature (p,q) and dimension 2^{p+q} is constructed through specific rules governing how the geometric product interacts with the



(a) **(b)**

Fig. 1: (a) 2D vector field (Optical Flow, left) and depth map (Monocular Depth Estimation, right) of a video frame from the COHFACE dataset. **(b)** PSDs of the averaged motion and depth signal components, compared to the ground truth PSD (dashed gold line), for the same COHFACE subject.

basis elements[1] $\{\mathbf{e}_i\}_{i=1}^{n}$ of $\mathbb{R}^n$, that is:

$$e_i^2 = 1 \ (1 \leq i \leq p), \qquad e_j^2 = -1 \ (p < j \leq n), \qquad e_i e_j = -e_j e_i, (i \neq j) \quad , \quad (3)$$

where $e_i e_j$ (also written e_{ij}) denotes the geometric product of e_i and e_j.

In this paper, we explore various Clifford algebra signatures: $Cl_{2,0}$ (s.t. $e_1^2 = e_2^2 = 1$), $Cl_{0,2}$ (s.t. $e_1^2 = e_2^2 = -1$) and $Cl_{3,0}$ (s.t. $e_1^2 = e_2^2 = e_3^2 = 1$), $Cl_{0,3}$ (s.t. $e_1^2 = e_2^2 = e_3^2 = -1$). $Cl_{2,0}$ and $Cl_{3,0}$ are generated from $\mathbb{R}^2$ and $\mathbb{R}^3$, respectively. $Cl_{2,0}$ includes $\mathbb{C}$ as its even sub-algebra, while $Cl_{3,0}$ contains two sub-algebras, one isomorphic to $\mathbb{C}$ (the center) and the other to the quaternion algebra $\mathbb{H}$ (the even sub-algebra). $Cl_{0,2}$ is isomorphic to $\mathbb{H}$, while $Cl_{0,3}$ comprises two distinct copies of $\mathbb{H}$.

In order to build Clifford neural architectures, the geometric product is taken in its matrix form. For instance, in a 4-dimensional Clifford algebra, the coefficients of the right product ba of two multivectors $a = a_0 + a_1 e_1 + a_2 e_2 + a_{12} e_{12}$ and $b = b_0 + b_1 e_1 + b_2 e_2 + b_{12} e_{12}$ are obtained as:

$$W_a^R \, \underline{b} = \begin{pmatrix} a_0 & \gamma_1 a_1 & \gamma_2 a_2 & -\gamma_1\gamma_2 a_{12} \\ a_1 & a_0 & -\gamma_2 a_{12} & \gamma_2 a_2 \\ a_2 & \gamma_1 a_{12} & a_0 & -\gamma_1 a_1 \\ a_{12} & a_2 & -a_1 & a_0 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_{12} \end{pmatrix} \quad , \quad (4)$$

where $\underline{b} \in \mathbb{R}^{2^{p+q}}$ represents the coefficients vector of $b \in Cl_{p,q}$, while $\gamma_1 = \gamma_2 = 1$ for $Cl_{2,0}$ and $\gamma_1 = \gamma_2 = -1$ for $Cl_{0,2}$.

Clifford neural layers This section introduces the Clifford neural layers which define learnable parameters $w \in Cl_{p,q}$ representing, through the action of the Clifford product, transformations of their hypercomplex inputs in $Cl_{p,q}$.

[1] To ensure consistent notation, we introduce three symbols for denoting the i-th basis element in $\mathbb{R}^n$. For example, in $\mathbb{R}^2$ and $Cl_{2,0}$, $\mathbf{e}_1 = (1,0)^T$ represents a vector in $\mathbb{R}^2$, $e_1 = 0 + 1e_1 + 0e_2 + 0e_{12}$ represents a multivector in $Cl_{2,0}$, and $\underline{e}_1 = (0,1,0,0)^T$ denotes the coefficients of e_1, i.e., a vector in $\mathbb{R}^{2^{p+q}}$, where $p = 2, q = 0$.

Linear layer. A traditional linear layer is an application $l : \mathbb{R}^d \to \mathbb{R}^k$ followed by a component-wise non-linear activation function $\sigma : \mathbb{R} \to \mathbb{R}$, such that its output amounts to:

$$z_{c'}^{(l)} = \sigma\left(\sum_{c=1}^{d} w_{c,c'}^{(l)}\, x_c\right), \qquad \forall c' = 1, \ldots, k \quad .$$

Essentially, the projection of the output onto the c'-th direction in the $\mathbb{R}^k$ space is determined by the inner product, computed in $\mathbb{R}^d$ between the c'-th learnable weight vector $\boldsymbol{w}_{c'} \in \mathbb{R}^d$ and the input $\boldsymbol{x} \in \mathbb{R}^d$.

Clifford linear layer. Unlike traditional linear layers and Clifford layer models [3, 44], the Clifford linear layer operates entirely within the Clifford algebra, $Cl_{p,q}$. Indeed, for an input $x \in Cl_{p,q}$ and a learnable parameter $w^{(l)} \in Cl_{p,q}$ it leverages the Clifford product as follows:

$$\underline{z}^{(l)} = W^R_{w^{(l)}}\, \underline{x} \quad , \tag{5}$$

where the matrix $W^R_{w^{(l)}}$ is defined in Eq. (4) and $\underline{x}, \underline{z}$ denote the coefficients of $x, z \in Cl_{p,q}$. Unlike the layers proposed in [44] we do not make use of non-linear activation functions. Moreover, the d input and k output channels are fixed at 1, thus avoiding feature expansion. Alternatively, channels (indexed in Eq. (6) by c at the input level and by j at the layer's output) can be used to reduce the spatial dimensionality. This facilitates downsampling operations without the need for additional parameters or pooling layers, being its outputs computed as:

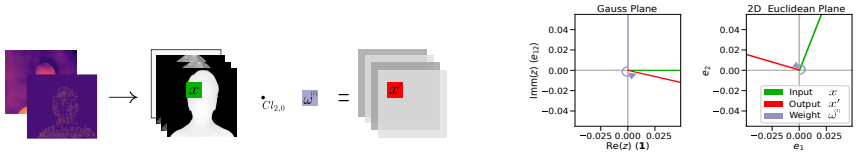
$$\underline{z}^{(l)}(j) = \sum_{c=1}^{HW} W^R_{w^{(l)}}(c, j)\, \underline{x}(c), \qquad \forall j = 1, \ldots, W'H' \quad . \tag{6}$$

Clifford 3D-convolutional layer. A Clifford 3D-convolutional layer is a mapping $l : Cl_{p,q}^{F\times W\times H} \to Cl_{p,q}^{F\times W\times H}$ that uses the Clifford product to convolve the input $\boldsymbol{x} \in Cl_{p,q}^{H\times W}$ with a single learnable element $\boldsymbol{w}^{(l)} \in Cl_{p,q}^{F\times H\times W}$:

$$\underline{z}^{(l)}(t, i, j) = \sum_{\tau=-\lfloor \frac{T}{2} \rfloor}^{\lfloor \frac{T}{2} \rfloor} \sum_{v,u=-\lfloor \frac{K}{2} \rfloor}^{+\lfloor \frac{K}{2} \rfloor} W^R_{w^{(l)}}(\tau, v, u)\, \underline{x}(t-\tau, i-v, j-u) \quad . \tag{7}$$

Similarly to Clifford linear layers in Eq. (5), we use Clifford 3D-convolutional layers without expanding channel dimensionality or applying non-linear activation functions.

The Geometric Bias Intuition. Clifford Algebra networks aim to learn transformations such as rotations, reflections, roto-translations, and symmetries. For example, multiplying a complex number by the imaginary unit i induces a 90°



(a) (b)

Fig. 2: **(a)** OF vector field and scalar MDE of a video frame processed through a Clifford linear layer. **(b)** Action of a learned algebra element on a single pixel, both modelled in $Cl_{2,0}$.

counterclockwise rotation in the Gaussian plane. To enable standard neural networks to learn such rotations, it would be intuitive to either represent both the input and weights in $\mathbb{C}$, or devise a method to apply i directly to vectors in $\mathbb{R}^2$. However, traditional neural networks operating in $\mathbb{R}^n$ can "only" project vectors onto output components via dot-product. Clifford Algebra addresses this limitation by integrating vectors and hypercomplex numbers with an operator (the Clifford product) ensuring that multiplying - say - a vector by a unit complex number yields a rotation in the plane. Similarly, rotations in $\mathbb{R}^3$ are represented by conjugating vectors with unit quaternions, while dual quaternions enable combined translation and rotation in higher-dimensional spaces [13]. To illustrate the geometric significance of the Clifford product in our practical scenario, consider modelling the input triplet $(\nu^{(i)}, \nu^{(j)}, z) \in Cl_{2,0}$, i.e. a pixel belonging to the recovered depth and motion maps (see Figure 2a). Since $Cl_{2,0} \simeq \mathbb{R}^2 \oplus \mathbb{C}$, we can represent $(\nu^{(i)}, \nu^{(j)}, z)$ as the algebraic element $x = z + \nu^{(i)} e_1 + \nu^{(j)} e_2$, composed of the complex number z (with null imaginary part) and the vector $\nu^{(i)}\mathbf{e}_1 + \nu^{(j)}\mathbf{e}_2$, as illustrated by the green directions in Figure 2b. Additionally, consider the element $\omega^{(l)} = -0.0258 - 0.0157e_1 + 0.0273e_2 + 0.0209e_{12}$ which represents a learned weight of a Clifford linear layer within a trained model designed to estimate the respiratory waveform from input x (*cfr.* 4). Multiplying x by the learned weight $\omega^{(l)}$, using Eq. (5), results in the rotated red directions depicted in Figure 2b. This example shows that any neural layer defining a learnable element $\omega^{(l)} \in Cl_{2,0}$ has the capability to learn rotations in both the Gaussian' and 2D Euclidean planes. Depending on the specific algebraic-geometric content within $Cl_{p,q}$, such example can be generalized to any geometric transformation in any hypercomplex space of arbitrary dimension.

4 Methods

4.1 Model Architecture

Building upon the model outlined in Eq. (2), the motion $x(t, i, j) \in \mathbb{R}^3$ associated to each pixel and estimated from an RGB video, can be interpreted as a triplet:

$$x(t, i, j) = (\nu^{(i)}, \nu^{(j)}, z)_{t,i,j} \quad . \tag{8}$$

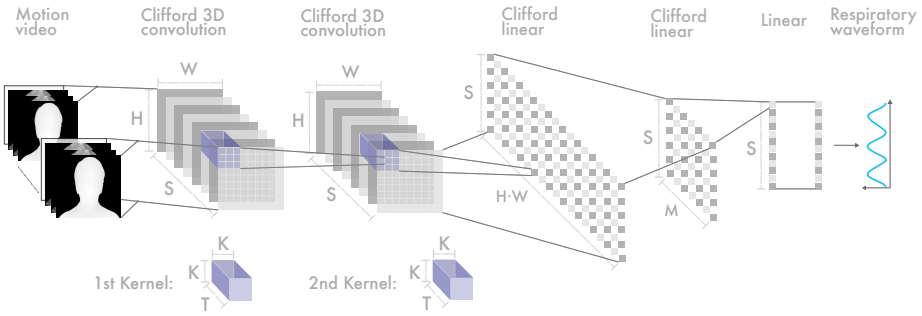


Fig. 3: `CliffPhys`$_{Cl_{p,q}}$ model architecture. The 3D motion video is taken as input. By learning Clifford transformations (i.e., elements of the Clifford Algebra of signature p, q), the network transforms the input data into the corresponding respiratory waveform. Parameters values: $S = 399$, $W = 36$, $H = 36$, $M = 128$, $K = 3$ and $T = 15$

Here, (i, j) denotes the spatial position of the pixel, while t refers to frame position within the video.

The uniqueness of the `CliffPhys` model family lies in the handling of the input motion video as an element $\boldsymbol{x} \in Cl_{p,q}^{S\times W\times H}$ of the Clifford Algebra $Cl_{p,q}$, with signature (p, q) as a hyperparameter. The input video is processed through five neural layers $l_0, \ldots, l_4$: two 3D-convolutional Clifford layers, two Clifford linear layers, and one traditional linear layer. `CliffPhys` model architecture is shown at a glance in Fig. 3.

Each 3D-convolutional Clifford layer l_κ ($\kappa = \{0, 1\}$), defines a kernel comprising $T \times K \times K$ learnable algebra elements, forming a filter bank with a single weight element $\boldsymbol{w}^{(l_\kappa)} \in Cl_{p,q}^{T\times K\times K}$. Setting $\boldsymbol{z}^{(l_0)} = \boldsymbol{x}$, these layers perform the following mapping, according to the 3D Clifford convolution:

$$l_\kappa : Cl_{p,q}^{S\times W\times H} \to Cl_{p,q}^{S\times W\times H}, \qquad \boldsymbol{z}^{(l_\kappa)} \mapsto \boldsymbol{z}^{(l_{\kappa+1})} \quad . \tag{9}$$

The Clifford linear layer l_2 is fed with flattened data from the previous layer. Temporal dimension is handled grouping the input in batches of size S. The spatial organization is then rearranged into a single pixel list, where $z^{(l_2)}(t, i, j)$ becomes $z_t^{(l_2)}(c)$, with $c = 1, \ldots, HW$. In the Clifford linear layers, l_κ, with $\kappa = \{2, 3\}$, the notion of channels is utilized for spatial downsampling as in Eq. (6). The first Clifford linear layer performs a mapping:

$$l_2 : Cl_{p,q}^{HW} \to Cl_{p,q}^{M}, \qquad \boldsymbol{z}_t^{(l_2)} \mapsto \boldsymbol{z}_t^{(l_3)} \quad . \tag{10}$$

The second linear layer l_3 completes the spatial down-sampling by implementing the mapping:

$$l_3 : Cl^M_{p,q} \to Cl_{p,q}, \qquad \boldsymbol{z}_t^{(l_3)} \mapsto z_t^{(l_4)} \quad . \tag{11}$$

Eventually, the traditional linear layer l_4 combines the coefficients $w_j^{(l_4)} \in \mathbb{R}$, $j = 1, \ldots, 2^{p+q}$ with the output of the previous layer, to produce the scalar $\hat{y}_t \in \mathbb{R}$, one for each frame. At this stage, the input is considered a vector $\underline{z}_t^{(l_4)} \in \mathbb{R}^{2^{p+q}}$ rather than an element of the algebra $z_t^{(l_4)} \in Cl_{p,q}$. The original temporal organisation is then recovered, yielding the prediction $\hat{\boldsymbol{y}} \in \mathbb{R}^S$ for the respiratory waveform.

5 Experiments

5.1 Datasets and Metrics

Training and experimental work are carried out on 3 datasets covering a diverse range of acquisition protocols (different ages, genders, or elicitation stimuli) and moving patterns (steady, moving heads, or talking):
SCAMPS [29] includes 2,800 synthetic videos (equivalent to 1.68M frames) with aligned cardiac and respiratory signals, and facial action intensities.
COHFACE [15] comprises 160 one-minute recordings from 40 subjects, featuring RGB videos, blood volume pulse (BVP) signals, and breathing waveforms.
BP4D+ [58] consists of 1400 multimodal recordings (140 subjects, 10 tasks designed to elicit different emotions). Data comprises 3d facial models, RGB and thermal videos, and physiological signals (including respiratory waveforms).

Model evaluation is conducted computing common metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Pearson Correlation Coefficient (PCC), and Concordance Correlation Coefficient (CCC).

5.2 Experimental Setup

Given the synthetic nature of the SCAMPS dataset, it has been employed only in the pre-training stage of the networks. COHFACE has been adopted for both training and testing, while BP4D + has been used solely for evaluation. Specifically, each $\texttt{CliffPhys}_{Cl_{p,q}}$ model has been pre-trained on SCAMPS and subsequently fine-tuned on the training set of the COHFACE dataset. The official training/validation/test splits, reported in [14], have been adopted.

Each input video is downsampled to 20 FPS and to a resolution of 36×36 pixels. The chosen spatial resolution balances image quality while suppressing camera noise [53]. Each video is then windowed into N non-overlapping windows of $S = 399$ frames (≈ 20 seconds) and finally standardized using the means and standard deviations computed on the COHFACE training set.

During the training phase, the respiratory ground truth (GT) $\boldsymbol{y} \in \mathbb{R}^F$ is filtered using a second-order Butterworth bandpass filter f_b with cutoff frequencies $[0.1, 0.5]$Hz, then resampled at 20Hz, and windowed into 399-sample windows.

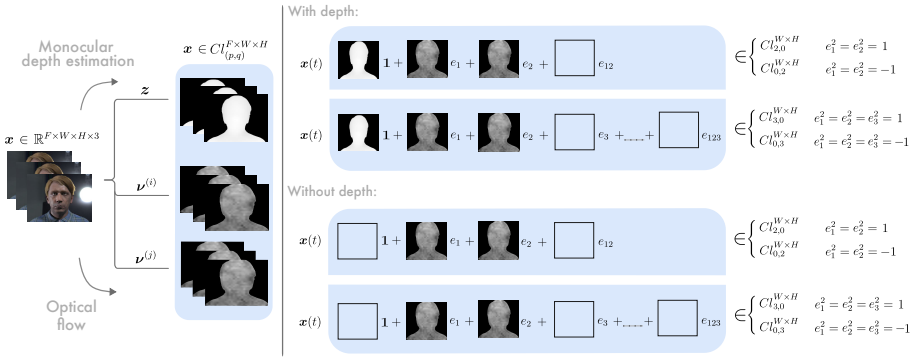


Fig. 4: Motion estimation. An input RGB video is converted into its 3D motion representation, i.e. a spatio-temporal organization of $Cl_{p,q}$ elements arranged into two configurations (with or without depth map). (p, q) is alternately set to $(2, 0)$, $(0, 2)$, $(3, 0)$, and $(0, 3)$, resulting in 8 possible models.

In the testing phase, each $\texttt{CliffPhys}_{Cl_{p,q}}$ model is evaluated by comparing the predicted respiration per minute (RPM) $\hat{h}$ with the ground truth RPM h. The predicted respiratory waveform is obtained by filtering (using f_b) the concatenated predictions related to the N video windows. Both $\hat{h}$ and h are computed by maximizing the PSD of the respective respiratory waveforms.

In order to assess the usefulness of depth information, each $\texttt{CliffPhys}_{Cl_{p,q}}$ model is trained and tested under two possible data configurations: with or without depth information. As qualitatively shown in Fig. 4, the depth-lacking configuration simply sets $z = 0$ in the triple $(\nu^{(i)}, \nu^{(j)}, z)_{t,i,j}$, for any t, i, j.

Each $\texttt{CliffPhys}_{Clp,q}$ model is trained with the AdamW optimizer using a weight decay of 0.0005, and a cyclically-updated learning rate (between 0.0001 and 0.001). Training utilizes the negative Pearson loss function [56], early stopping after 8 epochs of no improvement and batch size of 16 videos. Dropout is applied before each linear layer to mitigate overfitting. In the `CliffPhys` architecture, parameters are set as follows: $W = H = 36$, $S = 399$, $M = 128$, $T = 15$, and $K = 3$ (as described in Fig. 3). As mentioned in Sec. 4.1, the algebra signature p, q is a crucial model hyperparameter, determined through a grid search considering only theoretically supported values. Four models are considered: $Cl_{2,0}$, $Cl_{3,0}$, $Cl_{0,2}$, $Cl_{0,3}$. Depth maps are estimated using MiDaS DPT-Large with SwinV2 backbone [1, 39, 40]. Optical flow is computed using Raft-Small [30, 48].

5.3 Other models and baselines.

We evaluated $\texttt{CliffPhys}_{Clp,q}$ against nine approaches, including two motion-based baseline methods (`Median` and `Profile1D`) [52]. `Median` tracks each frame's median vertical optical flow over time. Conversely, the `Profile1D` approach estimates respiratory motion by correlating 1D image profiles [52].

We compared our methods to four recent neural techniques receiving raw RGB frames as input: `MTTS-CAN` [22], `BigSmall` [31], `ContrastPhys` [45] and `PhysFormer` [56]. The first two directly estimate the respiratory waveform from video, while the latter two are rPPG-based.

To evaluate the contribution of our hypercomplex architecture, we designed a neural model (`3D Conv`) that alternatively takes pre-processed frame representations $(\nu^{(i)}, \nu^{(j)}, z)$ or RGB data as input. The `3D Conv` model shares a similar architecture to $\texttt{CliffPhys}_{Clp,q}$, but employs conventional layers. Eventually, we asses the importance of the pre-processing phase by running the $\texttt{CliffPhys}_{Clp,q}$ on raw RGB frames.

5.4 Intra-Dataset Results

After pre-training and fine-tuning, $\texttt{CliffPhys}_{Cl_{p,q}}$ models' performance has been assessed on the COHFACE test set. Results are reported in Tab. 1. Notably, the

Table 1: Intra-dataset results. The $\texttt{CliffPhys}_{Cl_{p,q}}$ (and the 3D conv) models are pre-trained on SCAMPS and fine-tuned on the COHFACE training set. Testing is performed on the COHFACE test set. *Measured in $\frac{\text{breaths}}{\text{min}}$, **rPPG-based methods.

Input	Model	RMSE*↓	MAE* ↓	MAPE ↓	PCC↑	CCC ↑
$(\nu^{(i)}, \nu^{(j)}, z)$	$\texttt{CliffPhys}_{Cl_{0,2}}$	**1.97**	**0.83**	**10.62**	**0.86**	**0.83**
$(\nu^{(i)}, \nu^{(j)}, z)$	$\texttt{CliffPhys}_{Cl_{0,3}}$	2.30	0.99	12.23	0.80	0.78
$(\nu^{(i)}, \nu^{(j)}, z)$	$\texttt{CliffPhys}_{Cl_{2,0}}$	2.55	1.06	12.47	0.78	0.75
$(\nu^{(i)}, \nu^{(j)}, z)$	$\texttt{CliffPhys}_{Cl_{3,0}}$	2.27	0.99	12.28	0.81	0.79
$(\nu^{(i)}, \nu^{(j)})$	$\texttt{CliffPhys}_{Cl_{0,2}}$	<u>2.20</u>	<u>0.90</u>	<u>11.35</u>	<u>0.82</u>	<u>0.80</u>
$(\nu^{(i)}, \nu^{(j)})$	$\texttt{CliffPhys}_{Cl_{0,3}}$	2.31	1.03	12.55	0.81	0.78
$(\nu^{(i)}, \nu^{(j)})$	$\texttt{CliffPhys}_{Cl_{2,0}}$	2.26	0.95	12.03	0.81	0.79
$(\nu^{(i)}, \nu^{(j)})$	$\texttt{CliffPhys}_{Cl_{3,0}}$	2.21	0.93	11.83	0.81	0.79
(R, G, B)	$\texttt{CliffPhys}_{Cl_{3,0}}$	6.59	4.15	43.41	0.27	0.18
$(\nu^{(i)}, \nu^{(j)}, z)$	3D Conv	4.64	3.62	29.86	0.07	0.07
(R, G, B)	3D Conv	4.38	3.60	37.36	0.10	0.06
(R, G, B)	`MTTS-CAN`	4.84	2.83	25.92	0.41	0.38
(R, G, B)	`BigSmall`	6.98	5.58	55.95	0.12	0.06
(R, G, B)	`PhysFormer`**	4.70	3.72	36.51	0.03	0.02
(R, G, B)	`ContrastPhys`**	5.78	4.73	39.19	-0.03	-0.03
$\nu^{(j)}$	`Median`	3.35	1.99	20.19	0.72	0.64
(R, G, B)	`Profile1D`	6.58	4.81	48.22	0.22	0.15

$\texttt{CliffPhys}_{Cl_{0,2}}$ model, incorporating depth information, exhibits overall highest performance across all metrics compared to its depth-lacking or different signatures counterparts. The top-performing model achieves RMSE and MAE below 2 and 0.90 RPM, respectively. Furthermore, its MAPE, PCC and CCC

values (10.62, 0.86, and 0.83 respectively), sensibly outperform those delivered by other $\texttt{CliffPhys}_{Cl_{p,q}}$ models. Remarkably, all $\texttt{CliffPhys}_{Cl_{p,q}}$ models outperform the adopted baselines and recently proposed neural approaches relying on RGB videos. Notably, the traditional deep model mimicking the `CliffPhys`' architecture (3D Conv) underperforms all models. These results, witness the importance of: 1) providing strong inductive bias in the input data representation 2) incorporating depth information in the respiratory estimation process 3) opportunely exploit the geometric relationship between the recovered input representation.

5.5 Cross-Dataset Results

The cross-dataset evaluation is performed on the BP4D+ dataset. Some of the videos have been excluded from the analysis because they present an unusable ground-truth respiratory waveform due to artifacts and disturbances, as reported in [12]. Specifically, following the detailed procedure presented in [12], after band-pass filtering and normalization, signals with abnormal standard deviations in peak-to-peak and peak-trough measurements were discarded. Additionally, signals with duration below 30 seconds were removed, resulting in 212 genuine videos. The list of considered videos is reported in the Supplementary Material.

The results of BP4D + are reported in Tab. 2. In general, the considerations made for the intra-dataset case also apply in the cross-dataset scenario. Specifically, $\texttt{CliffPhys}_{Cl_{p,q}}$ models outperform all the baselines and RGB-based neural approaches, with models ingesting depth information (*cfr.* $\texttt{CliffPhys}_{Cl_{3,0}}$ and $\texttt{CliffPhys}_{Cl_{0,3}}$) providing sensibly better results according to all the adopted metrics.

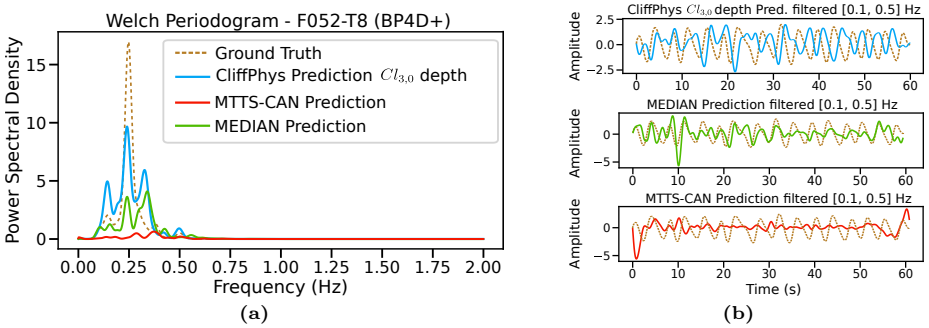


Fig. 5: Example of PSDs **(a)** and respiratory prediction signals **(b)** for subject F052-T8 in the BP4D+ dataset, displayed for MTTS-CAN (red), MEDIAN (green), and depth-informed $\texttt{CliffPhys}_{Cl_{3,0}}$ (light-blue), compared to the GT PSD (gold).

For a qualitative evaluation, we report in Fig. 5 the predicted respiratory waveform (Fig. 5b) and the PSD estimates (Fig. 5a) for the representative subject $F052$, trial $T8$ from the BP4D+ dataset. For comparison, results from the

Table 2: Cross-dataset results. The $\texttt{CliffPhys}_{Cl_{p,q}}$ are tested on the BP4D+ filtered dataset (212 videos). As in Tab. 1, models are pre-trained on SCAMPS and fine-tuned on the COHFACE training set. *Measured in $\frac{\text{breaths}}{\text{min}}$, **rPPG-based methods, ***the BIG-SMALL model is trained on BP4D+ following a multi-task learning approach.

Input	Algebra/Model	RMSE* ↓	MAE* ↓	MAPE ↓	PCC ↑	CCC ↑
$(\nu^{(i)}, \nu^{(j)}, z)$	$\texttt{CliffPhys}_{Cl_{0,2}}$	6.00	4.00	22.87	0.28	0.25
$(\nu^{(i)}, \nu^{(j)}, z)$	$\texttt{CliffPhys}_{Cl_{0,3}}$	<u>5.56</u>	<u>3.71</u>	20.68	<u>0.35</u>	<u>0.32</u>
$(\nu^{(i)}, \nu^{(j)}, z)$	$\texttt{CliffPhys}_{Cl_{2,0}}$	6.31	4.07	22.06	0.26	0.22
$(\nu^{(i)}, \nu^{(j)}, z)$	$\texttt{CliffPhys}_{Cl_{3,0}}$	**5.38**	**3.50**	**19.36**	**0.41**	**0.36**
$(\nu^{(i)}, \nu^{(j)})$	$\texttt{CliffPhys}_{Cl_{0,2}}$	5.75	3.86	21.04	0.34	0.30
$(\nu^{(i)}, \nu^{(j)})$	$\texttt{CliffPhys}_{Cl_{0,3}}$	6.03	4.16	22.16	0.34	0.27
$(\nu^{(i)}, \nu^{(j)})$	$\texttt{CliffPhys}_{Cl_{2,0}}$	5.84	3.84	<u>20.64</u>	0.32	0.27
$(\nu^{(i)}, \nu^{(j)})$	$\texttt{CliffPhys}_{Cl_{3,0}}$	6.33	4.23	22.91	0.21	0.18
(R, G, B)	$\texttt{CliffPhys}_{Cl_{3,0}}$	7.18	5.42	29.04	0.12	0.10
$(\nu^{(i)}, \nu^{(j)}, z)$	3D Conv	7.30	5.75	31.35	-0.09	-0.07
(R, G, B)	3D Conv	6.35	5.40	28.30	-0.03	-0.01
(R, G, B)	`MTTS-CAN`	6.86	5.73	31.11	0.20	0.14
(R, G, B)	`BigSmall`***	6.60	5.51	29.33	0.26	0.18
(R, G, B)	`PhysFormer`**	6.43	5.16	30.08	0.05	0.05
(R, G, B)	`ContrastPhys`**	5.40	3.82	26.53	-0.08	-0.04
$\nu^{(j)}$	`Median`	6.25	5.06	29.07	0.17	0.14
(R, G, B)	`Profile1D`	7.04	5.80	32.57	0.12	0.09

baseline `Median` approach and `MTTS-CAN` RGB-based neural method are depicted, too. As can be observed, $\texttt{CliffPhys}_{Cl_{3,0}}$ successfully estimates the subject's respiratory frequency, `Median` is able to partially recover the respiratory-related motion, however it cannot be considered the primary motion, if we inspect the PSD estimates. Similarly, the `MTTS-CAN` method struggles to identify a clear predominant motion, instead favoring a higher-frequency artifact.

6 Discussion

Experiments across intra- and cross-dataset settings demonstrate that $\texttt{CliffPhys}_{Cl_{p,q}}$ outperforms RGB-based deep learning approaches, 3D motion-based architectures, and baselines. This success likely stems from two primary reasons. Firstly, the data representation, motivated in Sec. 3.1, guides the network to prioritize spatio-temporal features associated with respiratory motion. Secondly, the $\texttt{CliffPhys}_{Cl_{p,q}}$ architecture is specifically designed to learn a series of geometrically-meaningful transformations preserving the relationships between different data channels throughout the prediction process. This geometric inductive bias is evident when comparing to results yielded by a standard convolutional architecture

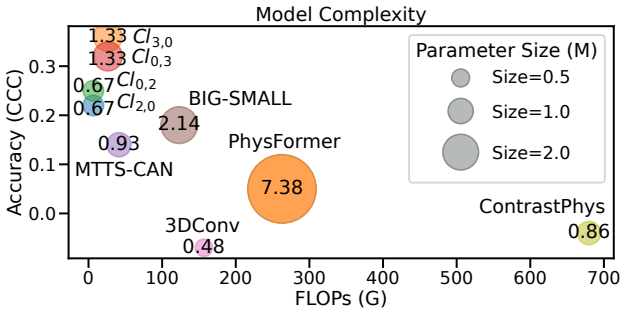


Fig. 6: Ball chart reporting the accuracy vs. computational complexity of the benchmarked models. The radius of each ball indicates model's size.

provided with the same input representation, which clearly struggle to properly address these relationships (cfr. Tab. 2).

In the intra-dataset setting, all $\texttt{CliffPhys}_{Cl_{p,q}}$ models perform similarly and outperform benchmark approaches. Overall best results are delivered by models ingesting both estimated apparent motion and depth maps. Among these, $\texttt{CliffPhys}_{Cl_{0,2}}$ and $\texttt{CliffPhys}_{Cl_{3,0}}$ achieve the highest performances. Notably, both algebras contain a copy of the real associative algebra $\mathbb{H}$ of quaternions. Unsurprisingly, a simpler algebra like $Cl_{2,0}$, closely tied to the group of rotations in the plane, behaves better when depth information is omitted.

Conversely, cross-dataset evaluation reveals that $\texttt{CliffPhys}_{Cl_{p,q}}$ models utilizing 8-dimensional Clifford algebras ($Cl_{3,0}$, $Cl_{0,3}$) outperform lower-dimensional models within the family. The overall stability of these models can be motivated by the fact the an 8-dimensional algebra has higher degrees of freedom. This allows a better handling of the richness of the input, represented by both depth estimation and optical flow. In turn, this could explain the relatively weaker performance of the depth-lacking version of $\texttt{CliffPhys}_{Cl_{3,0}}$.The scatter plot in Fig. 6 compares models effectiveness and efficiency. $\texttt{CliffPhys}_{Clp,q}$ models achieve superior CCC performance with notably lower parameter counts and floating-point operations per second (FLOPs).

7 Conclusion

This work presented the $\texttt{CliffPhys}_{Cl_{p,q}}$ model family, designed to automatically extract respiratory waveforms from 3D motion data estimated from RGB videos displaying human subjects. Extensive experiments, in both intra-dataset and cross-dataset settings, have shown that leveraging strong inductive biases in data representation yield excellent performance in estimating respiratory information. Notably, it appears evident that introducing depth information (albeit estimated from monocular RGB images) and exploiting the Clifford product to effectively transform the obtained representation, delivers the overall best results w.r.t. both baselines and most recent neural methods.

Acknowledgments

Financial support for this study was provided by a grant from Università degli Studi di Milano, Bando Linea 3 My First SEED, DM 737/2021 (MUR).

References

1. Birkl, R., Wofk, D., Müller, M.: Midas v3.1 – a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460 (2023) 10
2. Boccignone, G., D'Amelio, A., Ghezzi, O., Grossi, G., Lanzarotti, R.: An evaluation of non-contact photoplethysmography-based methods for remote respiratory rate estimation. Sensors **23**(7), 3387 (2023) 2
3. Brandstetter, J., Berg, R.v.d., Welling, M., Gupta, J.K.: Clifford neural layers for pde modeling. arXiv preprint arXiv:2209.04934 (2022) 3, 4, 6
4. Brieva, J., Ponce, H., Moya-Albor, E.: A contactless respiratory rate estimation method using a hermite magnification technique and convolutional neural networks. Applied Sciences **10**(2), 607 (2020) 2
5. Chen, J., Abbod, M., Shieh, J.S.: Pain and stress detection using wearable sensors and devices—a review. Sensors **21**(4), 1030 (2021) 1
6. Chen, W., McDuff, D.: Deepphys: Video-based physiological measurement using convolutional attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 349–365 (2018) 2
7. Chen, W.V., McDuff, D.J.: Deepphys: Video-based physiological measurement using convolutional attention networks. In: ECCV (2018) 2
8. Chiheb, T., Bilaniuk, O., Serdyuk, D., et al.: Deep complex networks. In: International Conference on Learning Representations (2017) 3
9. Danihelka, I., Wayne, G., Uria, B., Kalchbrenner, N., Graves, A.: Associative long short-term memory. In: International conference on machine learning. pp. 1986–1994. PMLR (2016) 3
10. Dorst, L., Fontijne, D., Mann, S.: Geometric Algebra for Computer Science: An Object-Oriented Approach to Geometry. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2009) 4
11. Fiedler, M.A., Rapczyński, M., Al-Hamadi, A.: Fusion-based approach for respiratory rate recognition from facial video images. IEEE Access **8**, 130036–130047 (2020) 2
12. Fiedler, M.A., Rapczyński, M., Al-Hamadi, A.: Fusion-based approach for respiratory rate recognition from facial video images. IEEE Access **8**, 130036–130047 (2020) 12
13. Grassucci, E., Mancini, G., Brignone, C., Uncini, A., Comminiello, D.: Dual quaternion ambisonics array for six-degree-of-freedom acoustic representation. Pattern Recognition Letters **166**, 24–30 (2023) 7
14. Heusch, G., Anjos, A., Marcel, S.: A reproducible study on remote heart rate measurement. CoRR **abs/1709.00962** (2017), http://arxiv.org/abs/1709.00962 2, 9
15. Heusch, G., Anjos, A., Marcel, S.: A reproducible study on remote heart rate measurement. arXiv preprint arXiv:1709.00962 (2017) 9
16. Hwang, H.S., Lee, E.C.: Non-contact respiration measurement method based on rgb camera using 1d convolutional neural networks. Sensors **21**(10), 3456 (2021) 2

17. Janssen, R., Wang, W., Moço, A., De Haan, G.: Video-based respiration monitoring with automatic region of interest detection. Physiological measurement **37**(1), 100 (2015) 2
18. Kempfle, J., Van Laerhoven, K.: Respiration rate estimation with depth cameras: An evaluation of parameters. In: Proceedings of the 5th international Workshop on Sensor-based Activity Recognition and Interaction. pp. 1–10 (2018) 4
19. Kumar, A.K., Ritam, M., Han, L., Guo, S., Chandra, R.: Deep learning for predicting respiratory rate from biosignals. Computers in Biology and Medicine **144**, 105338 (2022) 2
20. Lee, H., Lee, J., Kwon, Y., Kwon, J., Park, S., Sohn, R., Park, C.: Multitask siamese network for remote photoplethysmography and respiration estimation. Sensors **22**(14), 5101 (2022) 3
21. Lin, K.Y., Chen, D.Y., Tsai, W.J.: Image-based motion-tolerant remote respiratory rate evaluation. IEEE Sensors Journal **16**(9), 3263–3271 (2016) 2
22. Liu, X., Fromm, J., Patel, S., McDuff, D.: Multi-task temporal shift attention networks for on-device contactless vitals measurement. Advances in Neural Information Processing Systems **33**, 19400–19411 (2020) 2, 11
23. Liu, X., Hill, B., Jiang, Z., Patel, S., McDuff, D.: Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5008–5017 (2023) 2
24. Liu, X., Narayanswamy, G., Paruchuri, A., Zhang, X., Tang, J., Zhang, Y., Sengupta, R., Patel, S., Wang, Y., McDuff, D.: rppg-toolbox: Deep remote ppg toolbox. Advances in Neural Information Processing Systems **36** (2024) 1
25. Lounesto, P.: Clifford algebras and spinors. In: Clifford Algebras and Their Applications in Mathematical Physics, pp. 25–37. Springer (2001) 4
26. Luguern, D., Perche, S., Benezeth, Y., Moser, V., Dunbar, L.A., Braun, F., Lemkaddem, A., Nakamura, K., Gomez, R., Dubois, J.: An assessment of algorithms to estimate respiratory rate from the remote photoplethysmogram pp. 304–305 (2020) 2
27. Mateu-Mateus, M., Guede-Fernández, F., ángel García-González, M., Ramos-Castro, J.J., Fernández-Chimeno, M.: Camera-based method for respiratory rhythm extraction from a lateral perspective. IEEE access **8**, 154924–154939 (2020) 2
28. McDuff, D.: Camera measurement of physiological vital signs. ACM Computing Surveys **55**(9), 1–40 (2023) 1
29. McDuff, D., Wander, M., Liu, X., Hill, B., Hernandez, J., Lester, J., Baltrusaitis, T.: Scamps: Synthetics for camera measurement of physiological signals. Advances in Neural Information Processing Systems **35**, 3744–3757 (2022) 2, 9
30. Morimitsu, H.: Pytorch lightning optical flow. https://github.com/hmorimitsu/ptlflow (2021) 10
31. Narayanswamy, G., Liu, Y., Yang, Y., Ma, C., Liu, X., McDuff, D., Patel, S.: Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements. arXiv preprint arXiv:2303.11573 (2023) 3, 11
32. Neumann, D.A.: Kinesiology of the musculoskeletal system-e-book: foundations for rehabilitation. Elsevier Health Sciences (2016) 4
33. Niu, X., Yu, Z., Han, H., Li, X., Shan, S., Zhao, G.: Video-based remote physiological measurement via cross-verified feature disentangling. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 295–310. Springer (2020) 2

34. Oh, K., Shin, C.S., Kim, J., Yoo, S.K.: Level-set segmentation-based respiratory volume estimation using a depth camera. IEEE Journal of Biomedical and Health Informatics **23**(4), 1674–1682 (2018) 4
35. Parcollet, T., Ravanelli, M., Morchid, M., Linarès, G., Trabelsi, C., De Mori, R., Bengio, Y.: Quaternion recurrent neural networks. arXiv preprint arXiv:1806.04418 (2018) 3
36. Pepe, A., Lasenby, J., Buchholz, S.: Cgaposenet+ gcan: A geometric clifford algebra network for geometry-aware camera pose regression. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6593–6603 (2024) 3
37. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE transactions on biomedical engineering **58**(1), 7–11 (2010) 2
38. Prathosh, A., Praveena, P., Mestha, L.K., Bharadwaj, S.: Estimation of respiratory pattern from video using selective ensemble aggregation. IEEE Transactions on Signal Processing **65**(11), 2902–2916 (2017) 2
39. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ICCV (2021) 10
40. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3) (2022) 10
41. Ren, Y., Syrnyk, B., Avadhanam, N.: Dual attention network for heart rate and respiratory rate estimation. In: 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP). pp. 1–6. IEEE (2021) 3
42. Ren, Y., Syrnyk, B., Avadhanam, N.: Improving video-based heart rate and respiratory rate estimation via pulse-respiration quotient. In: Workshop on Healthcare AI and COVID-19. pp. 136–145. PMLR (2022) 3
43. Ruhe, D., Brandstetter, J., Forré, P.: Clifford group equivariant neural networks. arXiv preprint arXiv:2305.11141 (2023) 3
44. Ruhe, D., Gupta, J.K., De Keninck, S., Welling, M., Brandstetter, J.: Geometric clifford algebra networks. arXiv preprint arXiv:2302.06594 (2023) 3, 6
45. Sun, Z., Li, X.: Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In: European Conference on Computer Vision. pp. 492–510. Springer (2022) 1, 11
46. Suriani, N.S., Shahdan, N.S., Sahar, N.M., Taujuddin, N.S.A.M.: Non-contact facial based vital sign estimation using convolutional neural network approach. International Journal of Advanced Computer Science and Applications **13**(5) (2022) 2
47. Tay, Y., Zhang, A., Tuan, L.A., Rao, J., Zhang, S., Wang, S., Fu, J., Hui, S.C.: Lightweight and efficient neural natural language processing with quaternion networks. arXiv preprint arXiv:1906.04393 (2019) 3
48. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020) 10
49. Van Gastel, M., Stuijk, S., de Haan, G.: Robust respiration detection from remote photoplethysmography. Biomedical optics express **7**(12), 4941–4957 (2016) 2
50. Verkruysse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. Opt. Express **16**(26), 21434–21445 (2008). https://doi.org/10.1364/OE.16.021434 2
51. Wang, H., Zhou, Y., El Saddik, A.: Vitasi: A real-time contactless vital signs estimation system. Computers and Electrical Engineering **95**, 107392 (2021) 2, 3

52. Wang, W., den Brinker, A.C.: Algorithmic insights of camera-based respiratory motion extraction. Physiological measurement **43**(7), 075004 (2022) 2, 10
53. Wang, W., Stuijk, S., de Haan, G.: Exploiting spatial redundancy of image sensor for motion robust rppg. IEEE Transactions on Biomedical Engineering **62**(2), 415–425 (2015). https://doi.org/10.1109/TBME.2014.2356291 9
54. Yang, Y., Liu, X., Wu, J., Borac, S., Katabi, D., Poh, M.Z., McDuff, D.: Simper: Simple self-supervised learning of periodic targets. arXiv preprint arXiv:2210.03115 (2022) 1
55. Yu, Z., Shen, Y., Shi, J., Zhao, H., Cui, Y., Zhang, J., Torr, P., Zhao, G.: Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. International Journal of Computer Vision **131**(6), 1307–1330 (2023) 3
56. Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P., Zhao, G.: Physformer: Facial video-based physiological measurement with temporal difference transformer. arXiv preprint arXiv:2111.12082 (2021) 3, 10, 11
57. Zhang, A., Tay, Y., Zhang, S., Chan, A., Luu, A.T., Hui, S.C., Fu, J.: Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with $1/n$ parameters. arXiv preprint arXiv:2102.08597 (2021) 3
58. Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., et al.: Multimodal spontaneous emotion corpus for human behavior analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3438–3446 (2016) 2, 9