This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Overcoming Small Data Limitations in Video-Based Infant Respiration Estimation

Liyang Song†, Hardik Bishnoi†, Sai Kumar Reddy Manne,
Sarah Ostadabbas, Briana J. Taylor, Michael Wan‡

Northeastern University

†*equal contribution* ‡mi.wan@northeastern.edu

Abstract

The development of contactless respiration monitoring for infants could enable advances in the early detection and treatment of breathing irregularities, which are associated with neurodevelopmental impairments and conditions like sudden infant death syndrome (SIDS). But while respiration estimation for adults is supported by a robust ecosystem of computer vision algorithms and video datasets, only one small public video dataset with annotated respiration data for infant subjects exists, and there are no reproducible algorithms which are effective for infants. We introduce the ***a****nnotated* ***i****nfant* ***r****espiration dataset of* ***400*** *videos (AIR-400), contributing 275 new, carefully annotated videos from 10 recruited subjects to the public corpus. We develop the first reproducible pipelines for infant respiration estimation, based on infant-specific region-of-interest detection and spatiotemporal neural processing enhanced by optical flow inputs. We establish, through comprehensive experiments, the first reproducible benchmarks for the state-of-the-art in vision-based infant respiration estimation. We make our dataset, code repository, and trained models available for public use.*

1. Introduction

In infancy, healthy respiration is critical for optimal brain development and early detection and treatment of irregular respiration is critical to prevent neurodevelopmental impairment [14]. Breathing irregularities such as apnea or respiratory distress are significantly more common in high-risk neonates who often require post-delivery monitoring in neonatal intensive care units (NICUs) [6, 28]. However, respiration monitoring the NICU relies on contact sensors which can damage and cause pain to the sensitive skin of a neonate [2]. Moreover, gold standard respiration monitoring cannot be easily conducted in the home, post-discharge, despite substantial evidence linking apnea and sudden infant death syndrome (SIDS) [24]. The rise of human- and health-centric computer vision algorithms and widespread adoption of computationally-enabled baby monitors presents the opportunity to develop a contactless alternative to infant respiratory detection in home and medical settings. Recent advances in computer vision have included deep learning algorithms designed and trained to track infant-specific poses, states, and actions [4, 12, 13, 36]. Progress in infant respiration waveform and rate estimation, however, has been limited by the scarcity of annotated video data [26, 27]. As we will explain, most existing datasets are private and cannot be easily used or validated, and the sole work featuring a public dataset and algorithm for this task faces major limitations due to issues with dataset size and model reproducibility.

In this paper, we establish infant respiration estimation as a reproducible and benchmarkable task in computer vision for the first time, through three key contributions. First, we introduce a new video dataset with carefully synchronized and annotated respiration waveforms, which triples the amount of publicly accessible data, in terms of both the number of clips and the diversity of recruited infant subjects. Second, we propose a novel method for infant respiration waveform estimation that leverages infant pose detection and fine-grained region-of-interest (ROI) tracking, as illustrated in Figure 1. Third, we perform the most comprehensive evaluation to date of infant respiration estimation pipelines, systematically comparing optical flow and spatiotemporal network architectures, and rigorously testing reproducibility and generalizability across experimental settings. Together, these contributions provide a reliable foundation for advancing infant respiration monitoring within the computer vision community. Our dataset, code repository, and trained models can be found at https://github.com/michaelwwan/air-400.

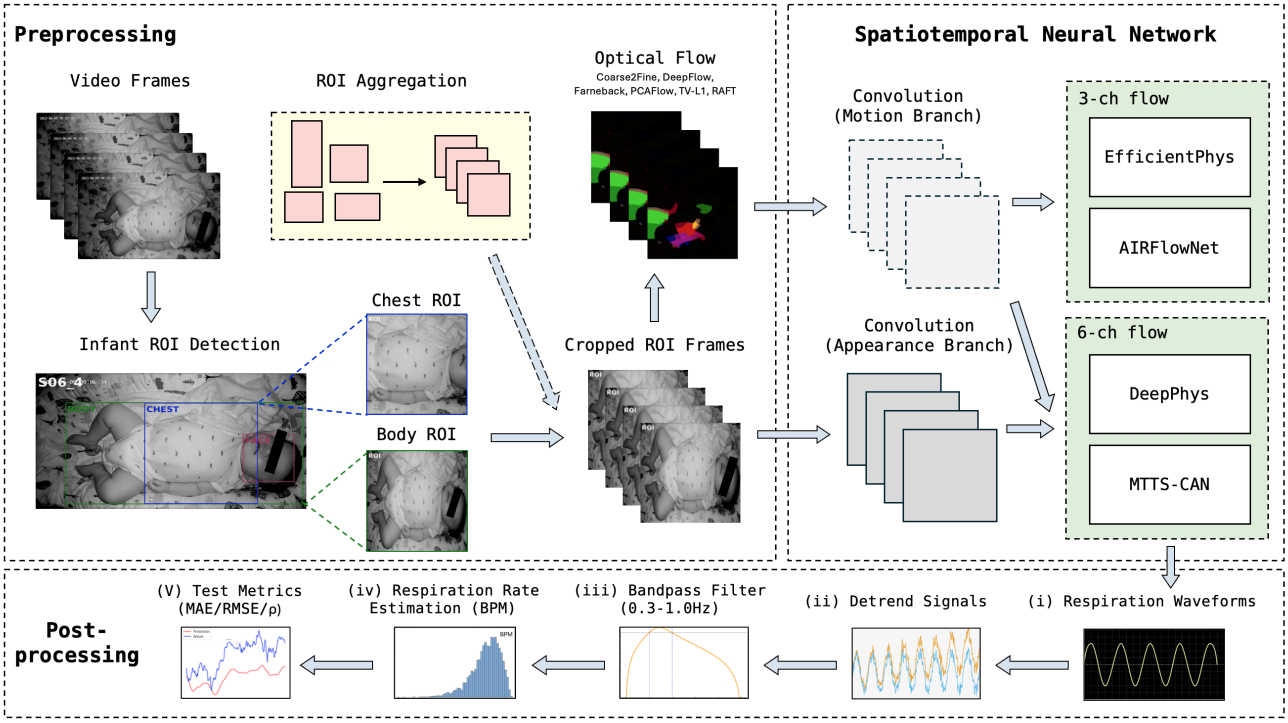

Figure 1. **Our infant respiration estimation pipeline** combines a preprocessing module with infant-specific spatial region-of-interest (ROI) detectors, spatiotemporal inference from neural networks, and signals post-processing. We implement two region-of-interest (ROI) strategies, an infant **body ROI** detector, and infant **chest ROI** detector. Clips are cropped to the ROI and processed by optical flow (Coarse2Fine [19], DeepFlow [37], Farnebäck [8], PCAFlow [38], TV-L1 [31, 40], RAFT [32]). Spatiotemporal neural networks (DeepPhys [3], MTTS-CAN [20], EfficientPhys [21], AIRFlowNet [23]) then process appearance and motion inputs, and post-processing via detrending and bandpass filtering is applied to obtain respiration rates and waveforms.

2. Related Work

Video-based physiological waveform and rate estimation has advanced rapidly. One of the most successful early models, DeepPhys [3], estimated both the cardiovascular and respiratory amplitude waveform by tracking changes in skin reflectance using a convolutional neural network with an attention mechanism to analyze differences in nearby frames. Subsequent methods found success applying 3D convolutional neural networks [39], though at a higher computational cost. Newer models such as the multi-task temporal shift convolutional attention network (MTTS-CAN) [20] and EfficientPhys [21] reduce the computational burden by replacing 3D convolutional networks with temporal shift modules combined with 2D convolutional processing.

Alongside methodological advances, vision-based physiological estimation has benefited from the public release of video datasets featuring a large number of adult subjects, accompanied by carefully calibrated ground truth physiological signals. Most notably, these include the COHFACE dataset [11], the MAHNOB dataset [30], and the synthetic SCAMPS [25] dataset. Characteristics of these datasets are tabulated alongside public and private datasets featuring both adult and infant subjects in Table 1. Focusing on respiration in the adult domain, there are 687 ground-truth videos from 67 real subjects publicly available with ground truth respiration signals, in addition to 2,800 videos from 2,800 synthetic subjects from SCAMPS.

By contrast, infant respiration estimation has been constrained by severe data scarcity, due to the complexity of collecting infant video data with corresponding ground truth respiration waveforms. The only public, annotated video data on infant respiration comes from our previous work in Manne et al. [23], which introduced the **a**nnotated **i**nfant **r**espiration dataset of **125** videos, or **AIR-125**, consisting of 125 short video clips drawn from eight infant subjects. This provided just enough data to train and test a preliminary infant respiration estimation model. Unfortunately, our more recent and comprehensive analysis documented in Section 5 shows that the strong numerical results obtained in [23] do not generalize to other choices of train–test split in the AIR-125 dataset, and are also not reproducible in exhaustive cross-validation testing in our newly expanded dataset. In this paper, we aim to rectify issues of data scarcity and model reproducibility with our updated dataset, methodology, and experimental contributions.

Table 1. **Datasets available to support video-based human respiration estimation**, highlighting the extreme scarcity of video data available featuring infant subjects. PPG: photoplethysmogram, PR: pulse rate, RR: respiratory rate, Resp: respiration waveform, AU: action unit, BVP: blood volume pulse, EEG: electroencephalogram, ECG: electrocardiogram, SpO2: blood oxygenation.

Dataset	Ground Truth	Domain	Videos	Subjects	Public
SCAMPS [25]	PPG, PR, RR, Resp, AU	Adult	2,800	2,800	✓
COHFACE [11]	Resp, BVP	Adult	160	40	✓
MAHNOB [30]	ECG, EEG, Resp	Adult	527	27	✓
AFRL [7]	ECG, EEG, PPG, PR, RR	Adult	300	25	✗
OBF [18]	RR, PPG, ECG	Adult	212	106	✗
Villarroel *et al.* [34]	Resp, PPG, SpO2	Infant	384	30	✗
Földesy *et al.* [10]	Resp	Infant	1,440	7	✗
Kyrollos *et al.* [17]	Resp	Infant	20	1	✗
Lorato *et al.* [22]	Resp	Infant	90	2	✗
Tveit *et al.* [33]	RR	Infant	6	2	✗
AIR-125 [23]	Resp, RR, Pose	Infant	125	8	✓
AIR-400 ***(Ours)***	Resp, RR, Pose	Infant	400	18	✓

3. An Expanded Infant Respiration Dataset

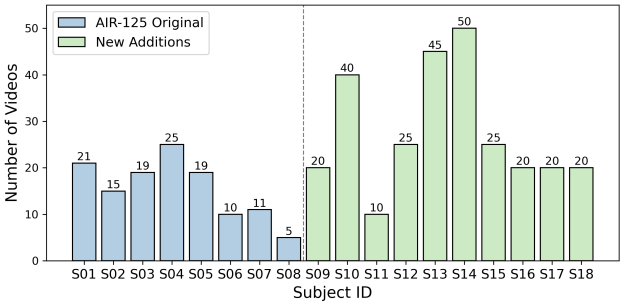

Figure 2. **Number of videos per subject in AIR-400.** The dataset comprises 400 videos from 18 infant subjects. Blue bars (S01–S08) represent the AIR-125 dataset [23] (125 videos from 8 subjects), while green bars (S09–S18) show our additions (275 videos from 10 new subjects from the same study as in [23]). Each bar indicates the number of 60 s clips available for that subject.

We expand our previous dataset of 125 infant respiration videos from Manne et al. [23] with an additional 275 clips drawn from new subjects from the same study population, formatted and annotated by our interdisciplinary team in the same way. We package the datasets together to produce a unified **a**nnotated **i**nfant **r**espiration dataset of **400** videos, or **AIR-400**. We next describe our process for collecting, curating, and annotating the 275 new videos, and characterize the AIR-400 dataset as a whole. See Figure 2 for a breakdown of AIR-125 and AIR-400 by subject.

3.1. Data Collection and Curation

We collected ∼100 hours of overnight baby monitor footage from 10 more infant subjects from the same study as in [23], under Northeastern Institutional Review Board approval (IRB #22-11-32). Participants ranged from 4–11 months in age. We provided baby monitor cameras to caregivers, with instructions to mount them overhead for a whole-body view of the infant. Raw recordings were captured for hours-long sessions overnight, and infant sleep is punctuated by periods of arousal and waking, as well as caretaker interactions. Footage was captured in both RGB and infrared (IR) modes, automatically determined by the camera based on the availability of ambient lighting.

Technical constraints presently stand in the way of fully-automated pediatric health monitoring from long-form video footage. Extreme infant movements and postures and anomalous events such as caretaker interactions can hamper effective tracking. High computational costs of preprocessing and training also prevent rapid experimentation needed at the research stage. Thus, in order to provide a feasible starting point for infant respiration estimation, we follow AIR-125 and curate a set of 60 s videos, featuring relatively clean samples of infant respiration behavior.

We were guided by three primary criteria to ensure dataset quality and utility. First, we favored highly visible chest and abdomen movements to enable effective visual annotation of respiratory cycles. We avoided videos with obstructions such as blankets, pacifiers, or toys. Second, we excluded excessive subject movement, caretaker interactions, or camera shake, which obscure respiratory movements. This consideration was especially important for the IR footage, where the combination of subtle thoracic movements, lower frame rates, and increased temporal noise in underexposed environments can create ambiguity. Third, we favored supine (face-up) positions for optimal chest visibility, while also including other sleeping orientations for variety. We also ensured diversity in respiratory patterns by selecting videos across an extended range of infant breathing rates (15–40 breaths-per-minute (BPM)). This range encompasses both typical clinical norms (30–60 BPM) [9] and challenging real-world scenarios where measured rates appear lower due to movement artifacts or natural breathing pauses during infant sleep. See Figure 3 for an overview of composition of the resulting dataset.

3.2. Annotation Process

To derive the ground truth respiration waveforms, we used the VGG Image Annotator software [5] to systematically analyze respiratory cycles and manually mark the peaks of the respiration waveform throughout each video sequence. The annotation was performed by one of the authors through iterative video playback, with frequent pausing, frame-by-frame advancement, and backtracking to ensure accurate peak identification. Respiration peaks were defined as the exhalation onset point (the transition from inspiration to expiration) [35], identified by observing the reversal from chest expansion to contraction. In cases where this transition was ambiguous, the minimum chest or abdomen position (point of maximum compression) was used as an alternative marker [16, 29].

Identification of peaks relied on multiple visual indica-

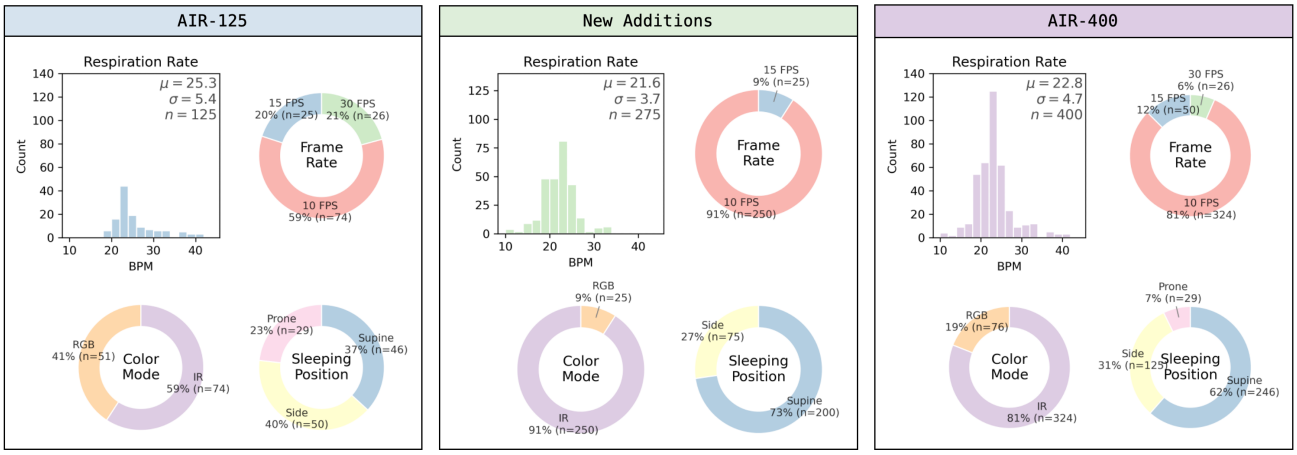

Figure 3. **Dataset statistics for AIR-125 [23], our newly contributed 275 videos, and the combined AIR-400.** Each panel shows a respiration-rate histogram, plus ring charts for frame rate, color mode (RGB vs infrared (IR)), and sleeping position (supine, side, and prone). The expanded dataset provides broader respiratory pattern diversity and varied recording conditions.

tors: direct thoracic and abdominal movements, facial cues (nostril and mouth changes), indirect movements propagated to other body regions (shoulders, limbs), and clothing deformation patterns which often amplified subtle respiratory motion. This multi-marker approach was essential given the variability in infant positioning and partial occlusions common in naturalistic recordings. Each video received at least two complete annotation passes: an initial marking of respiration peaks followed by validation for temporal consistency and physiological plausibility.

Following [23], we converted discrete peak annotations into continuous respiration waveforms, as follows. First, an impulse signal $I(t)$ was generated with unit impulses at annotated peak timestamps:

$$I(t) = \sum_i \delta(t - t_i), \tag{1}$$

where t_i is the timestamp of the ith annotated respiration peak. This impulse signal was then convolved with a Gaussian kernel to produce a smooth waveform:

$$w(t) = (I * G_\sigma)(t) = \sum_i G_\sigma(t - t_i), \tag{2}$$

where G_σ is a Gaussian kernel with standard deviation σ. We choose $\sigma = 4$ empirically to produce physiologically plausible waveforms while accommodating temporal uncertainties inherent in manual annotation.

3.3. Dataset Statistics

The AIR-400 dataset adds 275 additional annotated infant respiration videos to an existing public corpus of 125, providing 400 annotated clips in all, from 18 infant subjects, as illustrated in Figure 2. Figure 3 presents the distributional characteristics of both components. The respiration rate distributions show complementary coverage, with AIR-125 exhibiting a mean of 25.3 BPM ($\sigma = 5.4$) and the new additions centering at 21.6 BPM ($\sigma = 3.7$). This combined distribution spans a broader physiological range (approximately 15–40 BPM), as discussed in Section 3.2, capturing both normal tidal breathing and elevated respiratory patterns common in infant monitoring scenarios.

Notable differences in recording modalities reflect technical constraints inherent to infant monitoring. The predominance of infrared footage in the new additions (91.5% IR vs 8.5% RGB) corresponds to overnight recording conditions where ambient lighting necessitates IR capture. This modality difference directly influences frame rate characteristics: IR recordings are constrained to 10 frames per second (fps) due to sensor limitations, while RGB captures operate at higher frame rates (15–30 fps). The original AIR-125 dataset's mixed frame rate distribution (59.2% at 10 fps, 20.8% at 30 fps) reflects its more balanced RGB–IR composition (40.8% RGB, 59.2% IR). This diversity in temporal resolution enables researchers to evaluate algorithm robustness across the heterogeneous capture conditions encountered in real-world deployment, where consumer baby monitors operate at varying frame rates depending on lighting conditions.

Sleeping position distributions differ between the collections, with the new additions containing no prone position recordings. This reflects a methodological choice in our curation criteria (Section 3.1), as prone positions occlude thoracic and abdominal movement markers, compromising annotation reliability. The respiratory signal in prone infants manifests primarily through subtle spinal and shoulder blade movements, which were challenging to annotate consistently compared to the clear thoracoabdominal excursions visible in supine (74.6%) and side (25.4%) positions.

While the inclusion of prone positions (23.2%) in AIR-125, alongside supine (36.8%) and side (40%) positions, increases variety at the cost of annotation reliability. We opted to focus on reliability, in line with our goal of establishing the first robust benchmark and reproducible method for the infant respiration estimation task.

4. Methodology

Our infant respiration estimation pipeline, illustrated in Figure 1, combines a preprocessing module with infant-specific spatial region-of-interest (ROI) detection with spatial neural network processing. We work in particular with an infant-specific body ROI detector, together with a novel face-aware chest ROI detector, designed to exploit the pose and camera geometry of baby monitors to isolate the infant chest area. The crop from our infant ROI detector is then enhanced with deep, fine-grained optical flow, and processed by a spatiotemporal neural network optimized to track physiological signals. Post-processing is applied to convert the output respiration waveform into a respiration rate estimation for evaluation.

4.1. Preprocessing: ROI and Optical Flow

Visual evidence of the respiratory process is most prominent in the chest area, but existing techniques for identifying the chest ROI fall short in the infant domain, primarily due to poor performance of face detection for infant subjects in-crib. We test two variations on an infant-specific method for identifying and cropping to the relevant area in the video feed, **body ROI** detector and **chest ROI** detector, defined as follows, and illustrated in Figure 4.

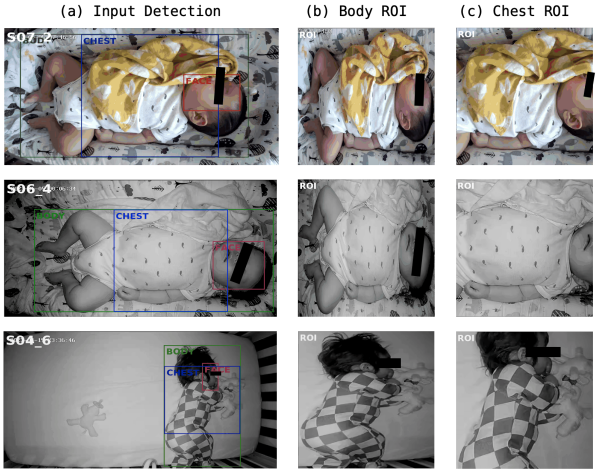

Figure 4. **Infant body region-of-interest (ROI) and chest ROI detections** across three subjects.

For each video frame $X_t \in \mathbb{R}^{H\times W\times 3}$, we apply YOLOv8 [15] to obtain the infant body bounding box $B_t = [x_t^B, y_t^B, w_t^B, h_t^B]$, and YOLOv8-Face [15], which is a lightweight model finetuned on human faces, to detect the infant face $F_t = [x_t^F, y_t^F, w_t^F, h_t^F]$. To mitigate infant face detection failures due to pose angle variation [36], face detection is applied to X_t and its $90°/180°/270°$ rotations, and the most confident prediction selected. Detections are performed every s frames (default $s = 10$) and at the last frame. Let $\mathcal{T}$ denote the set of detection indices. (Trivial detections are dropped for subsequent aggregation, if none remain, we fall back to the full frame (for body ROI) or a centered square of side $\min(W, H)$ (for chest ROI). All crops are clamped to image bounds, with out-of-bounds regions zero-padded.)

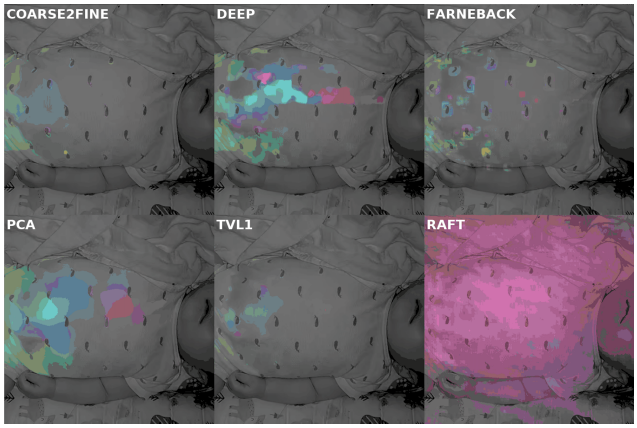

Figure 5. **Visualization of optical flow estimation on a sample frame from AIR-400.** Motion fields produced by six optical flow algorithms: Coarse2Fine, DeepFlow, Farnebäck, PCAFlow, TV-L1, RAFT.

Body ROI. If a valid person detection is present, the per-frame body ROI is B_t. If no person is detected (or a degenerate full-frame box is returned), we mark B_t as unreliable and fall back to the full frame for that time step. From the set $\{B_t\}_{t\in\mathcal{T}}$, we compute a single representative body ROI,

$$\tilde{B} = [\tilde{x}^B, \tilde{y}^B, \tilde{w}^B, \tilde{h}^B], \tag{3}$$

by aggregating centers with the median and sizes with a robust upper quantile:

$$\tilde{c}_x^B = \mathrm{median}_t\big(x_t^B + \tfrac{1}{2}w_t^B\big) \tag{4}$$

$$\tilde{c}_y^B = \mathrm{median}_t\big(y_t^B + \tfrac{1}{2}h_t^B\big) \tag{5}$$

$$\tilde{w}^B = \mathrm{P75}_t(w_t^B) \tag{6}$$

$$\tilde{h}^B = \mathrm{P75}_t(h_t^B), \tag{7}$$

and then recentering to $(\tilde{c}_x^B, \tilde{c}_y^B)$ and clamping to the image bounds. Optionally, we enlarge $\tilde{B}$ by a factor $k>1$ and reclamp.

Chest ROI. Given B_t and F_t, define their centers $b_t = \left(x_t^B + \frac{1}{2}w_t^B,\ y_t^B + \frac{1}{2}h_t^B\right)$ and $f_t = \left(x_t^F + \frac{1}{2}w_t^F,\ y_t^F + \frac{1}{2}h_t^F\right)$. We construct a face-aware chest center c_t^C by nudging the body center toward the face by a factor $\alpha \in (0, 1)$, but only along the body's longer axis to preserve anatomical plausibility:

$$c_t^C = \begin{cases} \left(b_{x,t} + \alpha(f_{x,t} - b_{x,t}),\ b_{y,t}\right) & \text{if } w_t^B \geq h_t^B, \\ \left(b_{x,t},\ b_{y,t} + \alpha(f_{y,t} - b_{y,t})\right) & \text{otherwise.} \end{cases} \tag{8}$$

The chest box side length is the body's short side, $s_t = \min(w_t^B, h_t^B)$, yielding the per-frame square $C_t = [x_t^C, y_t^C, s_t, s_t]$ centered at c_t^C and constrained to lie inside B_t and the image. If F_t is missing, we substitute $f_t \leftarrow b_t$ (no nudge). Aggregation mirrors the body ROI: median center, 75th-percentile side, recentering, clamp, and optional enlargement, producing $\tilde{C} = [\tilde{x}^C, \tilde{y}^C, \tilde{s}, \tilde{s}]$.

Optical Flow. We compute dense motion within the crop using one of: Coarse2Fine [19], DeepFlow [37], Farnebäck [8], PCAFlow [38], TV-L1 [31, 40], or RAFT [32], all illustrated in Figure 5. If desired, frames are resampled to a target rate f_s' before flow. The resulting flow $U \in \mathbb{R}^{T \times h \times w \times 3}$ (channels: u, v, magnitude) is normalized alongside the frames; if flow is disabled, we use 3-channel frame-to-frame RGB differences. Finally, we concatenate channels to form a 6-channel tensor per frame,

$$Z_t = \begin{cases} \left[u_t, v_t, \sqrt{u_t^2 + v_t^2}\right] & \text{(3-ch flow)}, \\ \left[U_t \,\|\, X_t^{\text{ROI}}\right] & \text{(6-ch motion} \,\|\, \text{appear.)}, \\ \left[\Delta X_t^{\text{ROI}} \,\|\, X_t^{\text{ROI}}\right] & \text{(no-flow ablation)}. \end{cases} \tag{9}$$

and uniformly chunk $Z_{1:T}$ for training or validation.

Body ROI uses a robust, rectangular crop centered on person detection; chest ROI is a square, face-aware crop anchored to the torso and gently biased toward the head. Both are temporally stabilized via median centers and upper-quantile sizes. This design enables a fair direct comparison of ROI choices under identical flow and model settings.

4.2. Spatiotemporal Neural Network

The cropped optical flow stream is passed into a suite of spatiotemporal neural networks designed for physiological signal processing. We experiment with a suite of models, including DeepPhys [3], MTTS-CAN [20], EfficientPhys [21], and the model by Manne et al. designed for infant respiration, AIRFlowNet [23]. Following the original dual-stream design, AIRFlowNet and EfficientPhys use 3-channel flow tensor comprising u, v, and magnitude $m = \sqrt{u^2 + v^2}$. DeepPhys and MTTS-CAN receive a 6-channel tensor per frame composed of both motion branch (flow or RGB-difference) and appearance branch (raw frames).

From the cropped clip $X_{1:T}^{\text{ROI}}$ and (optionally) its dense flow $U_{1:T}$, we construct framewise inputs $Z_t \in \mathbb{R}^{h \times w \times C}$ (defined in Section 4.1). Models output a framewise waveform $\hat{y}_{1:T}$ that approximates the ground truth respiratory waveform $y_{1:T}$. Because infant recordings exhibit heterogeneous durations and frame rates, training and validation are computed at the *subject* level. Mini-batch chunks are re-assembled per subject s to form contiguous signals $\hat{y}_{1:T_s}^{(s)}$ and $y_{1:T_s}^{(s)}$ before the loss is applied. This stabilizes optimization in the presence of variable-length clips and yields unbiased validation across subjects.

Loss. We adopt a frequency-domain objective that compares *normalized* power spectral densities (PSD) of prediction and target within a physiologically plausible band. For zero-mean signals, let $P_{\hat{y}}$ and P_y denote their PSDs (via real fast Fourier transform). With $\tilde{P}(k) = \dfrac{P(k)}{\sum_{j \in \mathcal{K}} P(j) + \varepsilon}$ over bins $\mathcal{K}$ inside the band (with small $\varepsilon > 0$ for numerical stability), the loss is

$$\mathcal{L}_{\text{PSD}} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \left(\tilde{P}_{\hat{y}}(k) - \tilde{P}_y(k)\right)^2. \tag{10}$$

We average $\mathcal{L}_{\text{PSD}}$ over subjects in the batch. This emphasizes correct spectral concentration around the respiratory fundamental while being robust to small shifts in phase and amplitude.

4.3. Post-Processing

At inference we aggregate all chunks per subject and apply a standardized pipeline to obtain BPM. We first undo differencing if applied in preprocessing, and then remove low-frequency drift with a second-order smoother:

$$\hat{x} = \left(I + (\lambda^2 D^\top D)\right)^{-1} x, \quad x_{\text{detrend}} = x - \hat{x}, \tag{11}$$

where D is the second-order difference operator and λ=100 by default. We apply a first-order Butterworth band-pass with subject-specific bounds (0.3–1.0 Hz, or 18–60 BPM). Filtering uses zero-phase `filtfilt` to avoid lag. Finally, we estimate the rate by detecting peaks of the filtered waveform, computing the mean inter-peak interval Δt, and returning $60/\Delta t$ BPM.

5. Evaluation and Results

We provide a comprehensive evaluation of various elements of the infant respiration estimation pipeline, including the choice of optical flow method, spatiotemporal network, ROI detection method, and the effect of training set size. We conducted experiments on our AIR-400 dataset, with 400 videos from 18 infant subjects. Spatiotemporal network and training configurations were standard and are detailed in the Supplementary Material. To robustly assess model performance, we performed all tests under six-fold, subject-wise cross-validation. In each fold, subjects were split randomly

into training (12 subjects), validation (3), and testing (3), ensuring disjoint sets. All reported metrics are averaged across the six folds. Evaluation metrics included mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation (ρ) between predicted and ground truth respiration signals.

5.1. Comprehensive Evaluation

We start with a evaluation of pipelines, pairing the different optical flow methods and spatiotemporal neural networks described in Section 4. Table 2 aggregates these results, all under six-fold cross-validation on AIR-400. Since the relative performance of these pipelines is not strongly affected by the ROI method, we highlight the reliable body-ROI results here and return to ROI comparisons below.

Our findings show that, outside of anomalous results under RAFT optical flow (discussed in the Supplementary Material), most models perform fairly capably, with many models achieving strong MAE scores near or below 4, and Pearson correlation (ρ) over 0.5. This falls short of the performance of the best models on adult respiration estimation, which achieve MAEs in the 1–2 range and ρs of 0.5–0.8 on the COHFACE dataset, but the adult subjects are stationary and training and testing is supported by more precise, sensor-based ground truth signals. Older spatiotemporal models such as DeepPhys and MTTS-CAN perform competitively with newer models, suggesting that the available data has not saturated the capabilities of the newer, more complex models. Despite the enlarged subject pool in AIR-400, the variation between folds in the cross validation is still relatively high, as shown in Figure 6. Thus, infant respiration estimation would likely be served by even larger datasets, especially in the absence of more precise ground truth respiratory signal capture.

5.2. Effect of Training Set Size

We further explore the small data constraints on infant respiration estimation. Figure 7 shows the effect of restricting the training set size, to various proportions by subject, across our suite of models. There is a clear overall trend of greater training data availability leading to stronger results, but the small dataset sizes leads to stochasticity that contravenes this effect in certain instances. Nonetheless, the benefit of increasing the dataset size does not appear to have saturated in this complex task. We do not specifically compare to training with AIR-125 here, due to the complexity of integrating its unevenly represented subjects into cross-validation on AIR-400, but the evaluation at the 25% training set size serves as an approximate substitute.

5.3. Comparison of ROI Methods

Table 3 offers a performance comparison of our body ROI and chest ROI detection methods in the respiration estimation pipeline. We test under six-fold cross-validation across a range of models, with the optical flow method fixed. Overall, the body ROI performs better than no ROI for three of the four architectures, whereas the more concentrated chest ROI yields mixed results, sometimes slightly degrading performance. Qualitatively, we find that the body ROI method provides the most reliable and stable view of infant respiration activity across subjects, but additional work is needed to translate this into consistent gains. The Supplementary Material discusses challenges with the chest ROI detection.

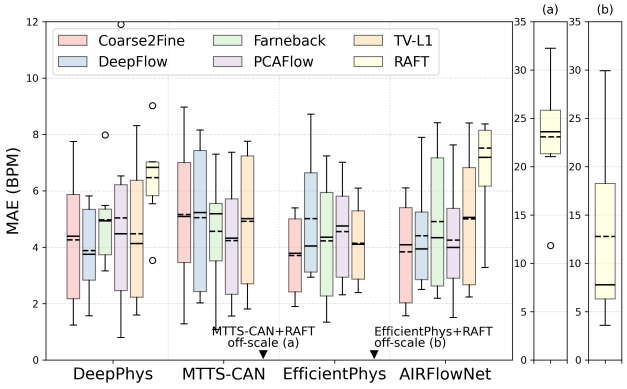

Figure 6. **Performance variation across cross-validation folds in our comprehensive evaluation** of infant respiration pipelines. Each box and whisker illustrates the MAE performance distribution in each of the six folds of AIR-400, with the box spanning the inter-quartile range (IQR) and whiskers extending to $1.5\times$IQR. The solid horizontal line inside each box denotes the median, and the dashed line denotes the mean.

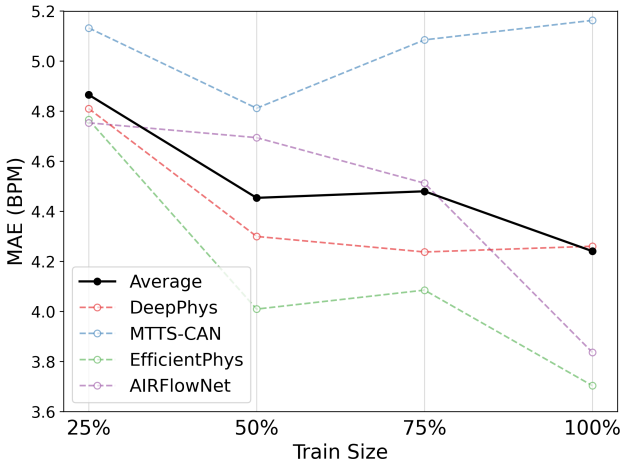

Figure 7. **Effect of training set size on respiration estimation.** Line charts show the performance of the DeepPhys, MTTS-CAN, EfficientPhys, and AIRFlowNet models—and their average—under 6-fold cross-validation, when training is restricted to random subsets of the indicated size. These results show, consistently across several models, the significant improvements enabled by the larger training set sizes from AIR-400.

Table 2. **Comprehensive evaluation of optical flow and spatiotemporal model pipelines,** under six-fold subject-wise cross-validation on AIR-400, with the infant body ROI detector enabled. Performance measured via mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation (ρ), each averaged across the six folds. Metrics are computed on post-processed respiration rates and averaged per subject within each fold. Best overall results by model in **bold**, and by optical flow (OF) method in *italics*. Across models, classical optical flow methods consistently outperformed RAFT.

OF \ Model	DeepPhys [3]			MTTS-CAN [20]			EfficientPhys [21]			AIRFlowNet [23]		
	MAE ↓	RMSE ↓	ρ ↑	MAE ↓	RMSE ↓	ρ ↑	MAE ↓	RMSE ↓	ρ ↑	MAE ↓	RMSE ↓	ρ ↑
Coarse2Fine [19]	4.26	6.82	0.45	5.16	8.30	0.28	***3.70***	***5.94***	***0.58***	**3.84**	**6.67**	**0.44**
DeepFlow [37]	***3.88***	***6.45***	***0.51***	5.05	8.21	0.36	5.01	7.54	0.34	4.41	7.56	0.39
Farnebäck [8]	4.97	8.28	0.27	4.56	7.72	0.44	*4.22*	*6.84*	*0.44*	4.91	8.00	0.38
PCAFlow [38]	5.04	7.69	0.41	***4.23***	***6.85***	***0.50***	4.55	7.11	0.42	4.25	6.95	0.47
TV-L1 [31, 40]	4.48	7.11	0.39	4.91	7.97	0.42	*4.14*	*6.59*	*0.47*	5.01	8.17	0.44
RAFT [32]	*6.47*	*8.36*	*0.35*	23.07	26.23	-0.17	12.77	14.57	0.10	7.52	9.85	0.26

Table 3. **Performance evaluation of ROI methods** across different models, under six-fold cross-validation on AIR-400, with Coarse2Fine optical flow processing. Strongest overall results per model in **bold**. Using the infant body ROI generally improves MAE and RMSE for DeepPhys, EfficientPhys, and AIRFlowNet, but not MTTS-CAN.

Model	ROI	MAE ↓	RMSE ↓	ρ ↑
DeepPhys	None	4.78	7.69	0.33
	Body ROI	**4.26**	**6.82**	**0.45**
	Chest ROI	4.45	6.92	0.39
MTTS-CAN	**None**	**4.35**	**7.59**	**0.38**
	Body ROI	5.16	8.30	0.28
	Chest ROI	4.90	7.67	0.38
EfficientPhys	None	4.57	7.08	0.39
	Body ROI	**3.70**	**5.94**	**0.58**
	Chest ROI	4.68	7.23	0.38
AIRFlowNet	None	4.76	7.79	0.32
	Body ROI	**3.84**	**6.67**	**0.44**
	Chest ROI	4.24	7.14	0.42

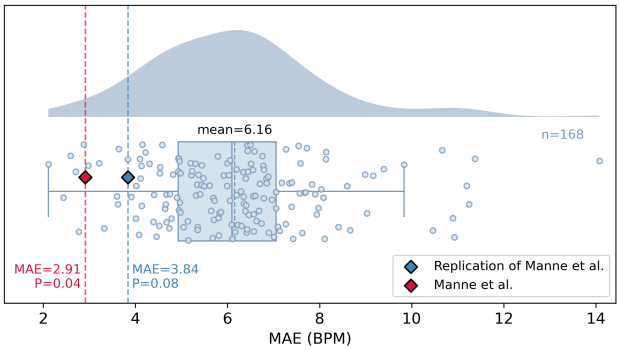

Figure 8. **Reproducing Manne et al.** Illustrating how reported and recreated results from [23] sit as outliers in the distribution of MAE results under all 168 train–test configurations in AIR-125 consistent with their documented specifications. We argue that our present, comprehensive cross-validation outcomes on AIR-400 represent more reproducible baselines for infant respiration estimation. The box plot shows the mean, median, and interquartile range (IQR) with whiskers at $1.5 \times$ IQR; the half-violin plot above shows a corresponding kernel density estimation.

5.4. Experimental Reproducibility and Variability

In Section 5.1, we established under comprehensive cross-validation that our best infant respiration models perform with MAEs of 3.7–4.0 BPM, and in Section 5.2, we showed that the larger training sets afforded by AIR-400 are essential for achieving this performance. These findings stand in apparent tension with results from our earlier infant respiration efforts in Manne et al. [23], where our best model, AIRFlowNet, achieved a MAE of 2.91 BPM, despite being trained on the smaller but similar AIR-125 dataset. We conducted a careful study to resolve this discrepancy, described in the Supplementary Materials. Here, we note our main findings: that (i) the precise results were not easily reproducible [23], but (ii) any results in the MAE 2.9–4.0 BPM range sit well below the mean of 6.16 BPM, corresponding to advantageous choices of subjects for the train–test split, and do not represent reliably achievable performance on AIR-125 (see Figure 8).

6. Conclusion & Outlook

In this work, we establish infant respiration estimation as a reproducible and benchmarkable task in computer vision for the first time, with the introduction of our AIR-400 dataset and our spatiotemporal estimation pipeline, as well as our comprehensive evaluation and resolution of reproducibility issues. Our results suggest that larger training datasets would support further performance gains, including perhaps by enabling estimation pipelines to make better use of region-of-interest detection and greater model capacities. Our study focuses on hand-curated clips featuring mostly-still infants with distinct breathing patterns, leaving room for future approaches to tackle more active subjects and noisier footage. In addition, downstream applications to the detection of breathing irregularities such as apnea or respiratory distress would be broadly impactful.

Acknowledgments

We gratefully acknowledge support for this research from a Northeastern University TIER 1 Seed Grant, and NSF CAREER Award #2143882.

References

[1] Sharib Ali, Christian Daul, and Walter Blondel. Robust and accurate optical flow estimation for weak texture and varying illumination conditions: Application to cystoscopy. In *2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2014. 11

[2] Oliver Bonner, Kathryn Beardsall, Nathan Crilly, and Joan Lasenby. 'there were more wires than him': the potential for wireless patient monitoring in neonatal intensive care. *BMJ innovations*, 3(1), 2017. 1

[3] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365, 2018. 2, 6, 8

[4] Amel Dechemi and Konstantinos Karydis. E-babynet: Enhanced action recognition of infant reaching in unconstrained environments. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:1679–1686, 2024. 1

[5] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, New York, NY, USA, 2019. ACM. 3

[6] Eric C Eichenwald, Kristi L Watterberg, Susan Aucott, William E Benitz, James J Cummings, Jay Goldsmith, Brenda B Poindexter, Karen Puopolo, Dan L Stewart, Kasper S Wang, et al. Apnea of prematurity. *Pediatrics*, 137(1), 2016. 1

[7] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 1462–1469. IEEE, 2014. 3

[8] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis (SCIA 2003)*, pages 363–370. Springer, 2003. 2, 6, 8

[9] Sina Fleming, Matthew Thompson, Richard Stevens, Carl Heneghan, Annette Plüddemann, Ian Maconochie, Lionel Tarassenko, and David Mant. Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. *The Lancet*, 377(9770):1011–1018, 2011. 3

[10] Péter Földesy, Ákos Zarándy, and Miklós Szabó. Reference free incremental deep learning model applied for camera-based respiration monitoring. *IEEE Sensors Journal*, 21(2): 2346–2352, 2020. 3

[11] Guillaume Heusch, André Anjos, and Sébastien Marcel. A reproducible study on remote heart rate measurement. *arXiv preprint arXiv:1709.00962*, 2017. 2, 3

[12] Xiaofei Huang, Nihang Fu, Shuangjun Liu, and Sarah Ostadabbas. Invariant representation learning for infant pose estimation with small data. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021. 1

[13] Xiaofei Huang, Lingfei Luan, Elaheh Hatamimajoumerd, Michael Wan, Pooria Daneshvar Kakhaki, Rita Obeid, and Sarah Ostadabbas. Posture-based Infant Action Recognition in the Wild with Very Limited Data. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4912–4921, Vancouver, BC, Canada, 2023. IEEE. 1

[14] Annie Janvier, May Khairy, Athanasios Kokkotis, Carole Cormier, Denise Messmer, and Keith J Barrington. Apnea is associated with neurodevelopmental impairment in very low birth weight infants. *Journal of perinatology*, 24(12): 763–768, 2004. 1

[15] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, 2023. 5

[16] S. Kohn, D. Waisman, J. Pesin, G. Huna-Baron, A. Faingersh, A. Landesberg, and A. Rotschild. Monitoring the respiratory rate by miniature motion sensors in premature infants: a comparative study. *Journal of Perinatology*, 36(4): 289–294, 2015. 3

[17] Daniel G Kyrollos, Joshua B Tanner, Kim Greenwood, JoAnn Harrold, and James R Green. Noncontact neonatal respiration rate estimation using machine vision. In *2021 IEEE Sensors Applications Symposium (SAS)*, pages 1–6. IEEE, 2021. 3

[18] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 242–249. IEEE, 2018. 3

[19] Ce Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. Doctoral thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2009. 2, 6, 8

[20] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020. 2, 6, 8

[21] Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5008–5017, 2023. 2, 6, 8

[22] Ilde Lorato, Sander Stuijk, Mohammed Meftah, Deedee Kommers, Peter Andriessen, Carola van Pul, and Gerard de Haan. Towards continuous camera-based respiration monitoring in infants. *Sensors*, 21(7):2268, 2021. 3

[23] Sai Kumar Reddy Manne, Shaotong Zhu, Sarah Ostadabbas, and Michael Wan. Automatic Infant Respiration Estimation from Video: A Deep Flow-Based Algorithm and a Novel Public Benchmark. In *Perinatal, Preterm and Paediatric Image Analysis*, pages 111–120. Springer Nature Switzerland,

Cham, 2023. Series Title: Lecture Notes in Computer Science. 2, 3, 4, 6, 8, 11

[24] Richard J Martin, Lisa J Mitchell, and Peter M MacFarlane. Apnea of prematurity and sudden infant death syndrome. *Handbook of clinical neurology*, 189:43–52, 2022. 1

[25] Daniel McDuff, Miah Wander, Xin Liu, Brian Hill, Javier Hernandez, Jonathan Lester, and Tadas Baltrusaitis. Scamps: Synthetics for camera measurement of physiological signals. *Advances in Neural Information Processing Systems*, 35: 3744–3757, 2022. 2, 3

[26] Sarah Ostadabbas, Christoph Bulach, David N Ku, Larry J Anderson, and Maysam Ghovanloo. A passive quantitative measurement of airway resistance using depth data. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5743–5747. IEEE, 2014. 1

[27] Sarah Ostadabbas, Nordine Sebkhi, Mingxi Zhang, Salman Rahim, Larry J Anderson, Frances Eun-Hyung Lee, and Maysam Ghovanloo. A vision-based respiration monitoring system for passive airway resistance estimation. *IEEE Transactions on biomedical engineering*, 63(9):1904–1913, 2015. 1

[28] Suzanne Reuter, Chuanpit Moser, and Michelle Baack. Respiratory Distress in the Newborn. *Pediatrics In Review*, 35 (10):417–429, 2014. 1

[29] J. X. Santos, P. Y. F. Silva, M. C. L. D. Cruz, B. F. V. E. Silva, I. G. Azevedo, and S. A. Pereira. Real-time changes in rib cage expansion and use of abdominal mechanical stimulation in newborns: a quasi-experimental study. *Revista Paulista de Pediatria*, 42:e2023032, 2023. 3

[30] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55, 2011. 2, 3

[31] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. TV-L1 Optical Flow Estimation. *Image Processing On Line*, 3:137–150, 2013. https://doi.org/10.5201/ipol.2013.26. 2, 6, 8, 11

[32] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. 2, 6, 8, 11

[33] Daniel Myklatun Tveit, Kjersti Engan, Ivar Austvoll, and Øyvind Meinich-Bache. Motion based detection of respiration rate in infants using video. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1225–1229. IEEE, 2016. 3

[34] Mauricio Villarroel, Sitthichok Chaichulee, João Jorge, Sara Davis, Gabrielle Green, Carlos Arteta, Andrew Zisserman, Kenny McCormick, Peter Watkinson, and Lionel Tarassenko. Non-contact physiological monitoring of preterm infants in the Neonatal Intensive Care Unit. *npj Digital Medicine*, 2(1):128, 2019. 3

[35] Diana Vitazkova, Erik Foltan, Helena Kosnacova, Michal Micjan, Martin Donoval, Anton Kuzma, Martin Kopani, and Erik Vavrinsky. Advances in respiratory monitoring: A comprehensive review of wearable and remote technologies. *Biosensors*, 14(2):90, 2024. 3

[36] Michael Wan, Shaotong Zhu, Lingfei Luan, Gulati Prateek, Xiaofei Huang, Rebecca Schwartz-Mette, Marie Hayes, Emily Zimmerman, and Sarah Ostadabbas. InfAnFace: Bridging the InfantâAdult Domain Gap in Facial Landmark Estimation in the Wild. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4486–4492, Montreal, QC, Canada, 2022. IEEE. 1, 5

[37] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow: Large displacement optical flow with deep matching. In *2013 IEEE International Conference on Computer Vision*, pages 1385–1392, 2013. 2, 6, 8

[38] Jonas Wulff and Michael J. Black. Efficient sparse-to-dense optical flow estimation using a learned basis and layers. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 120–130, 2015. 2, 6, 8

[39] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks, 2019. arXiv:1905.02419 [cs]. 2

[40] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l1 optical flow. In *DAGM-Symposium*, pages 214–223. Springer, 2007. 2, 6, 8, 11