RRPIPS: *R*espiratory Waveform *R*econstruction using Persistent Independent Particles tracking from video

Abstract

Non-contact monitoring videos capture subtle respiratory-induced motions, yet existing methods primarily focus on estimating respiratory rate (RR), neglecting the extraction of respiratory waveforms—a vital parameter that provides critical health information. We formulate video-based RR estimation as a Tracking All Points (TAP) problem and propose a coarse-to-fine, multi-frame *Persistent Independent Particle* (RRPIPs) framework for robust, multi-modal (RGB, NIR, IR) RR waveform estimation. Addressing the challenge of tracking minute, non-rigid pixel displacements caused by respiratory motions, our top-down approach magnifies respiratory motion using phase-based video magnification tuned to the respiratory frequency range and employs a pretrained RAFT optical flow model for initial region identification via a two-frame analysis. Coarse-scale tracking is performed using the RRPIPs model, while a Signal Quality Index (SQI) block evaluates the SNR of trajectories to refine high-respiratory-activity regions. These regions are upsampled, and fine-scale tracking is applied to extract precise waveforms. We curated a large-scale multimodal dataset for respiratory point tracking, combining *in-house* collected data and public datasets, with dense annotations of non-rigid pixel movements across multiple scales in key respiratory regions. Experimental results demonstrate that our framework achieves state-of-the-art accuracy (~1 MAE) and interpretability in respiratory waveform extraction across RGB, NIR, and IR modalities, effectively addressing multi-scale tracking and low-SNR challenges. Thorough ablation studies validate the contributions of each framework component, and we plan to open-source our codes and dataset to support further research.

Keywords

Particle Video, Contactless Respiratory rate, RR Tracking

ACM Reference Format:

. 2025. RRPIPS: *R*espiratory Waveform *R*econstruction using Persistent Independent Particles tracking from video. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1. Introduction

Respiratory rate (RR), or breathing frequency, measures the number of breaths per minute [3]. It is typically assessed by observing respiratory-induced movements in areas such as the chest, torso, shoulders, neck, and abdomen during inhalation and exhalation. RR measurements help classify respiratory patterns as normal (eupnea), excessively rapid (tachypnea), abnormally slow (bradypnea), or absent (apnea) respiration [11]. RR is a key marker for detecting conditions such as sleep apnea, sudden infant death syndrome, respiratory depression, and clinical deterioration [12, 13, 26, 50], as well as for diagnosing issues in post-anesthesia care, neonatal intensive care, and other clinical settings [25, 30]. Beyond clinical use, RR reflects stressors such as emotional stress, cognitive load, physical exertion, and fatigue [10, 20, 40, 47, 48, 60], driving advancements in automated measurement techniques [16, 40]. RR estimation methods are either contact-based [39], using sensors and equipment, or contactless [40], relying on body movements detected by video, laser, or radar, with contactless methods being more comfortable for prolonged use [41].

This study focuses on contactless respiration monitoring using video modalities (RGB, NIR, IR) available in devices like smartphones and computers, which capture respiratory-induced motions of visible organs. We address a critical research gap in non-contact respiratory rate (RR) estimation by focusing on the more complex problem of extracting respiratory waveforms, a graphical representation of changes in pressure, flow, and volume within the respiratory system. Beyond RR estimation, respiratory waveforms provide insights into breathing depth, timing, and consistency, aiding in the assessment of conditions such as apnea, abnormal respiration, spinal cord injury, diaphragmatic dysfunction, Cheyne-Stokes breathing patterns, central nervous system (CNS) changes, respiratory-metabolic imbalances, CO_2 levels [55], tidal volume [7, 65], breathing exercises and overall respiratory health.

While recent computer vision methods estimate RR by analyzing pixel movements, they fall short of reconstructing respiratory waveforms across diverse video modalities [41]. To bridge this gap, we propose formulating waveform estimation as a Tracking-All-Points (TAP) problem [14], enabling continuous tracking of subtle respiratory movements in videos. This task is inherently challenging due to low signal-to-noise ratios, small and non-rigid regions of interest, and significant variability across subjects, motion patterns, and backgrounds, with movements spanning only a few pixels [44, 52, 66]. Existing TAP-based deep learning models [22, 31, 63] underperform in such scenarios, especially on low-textured surfaces or under distribution shifts [15], highlighting the need for a specialized approach.

We propose a customized approach that leverages domain-specific priors to address respiratory waveform reconstruction through a multi-stage strategy. First, we localize relevant respiratory regions in video frames to minimize subsequent tracking computations. This initial step enables efficient and accurate region selection. by leveraging video motion magnification and utilizing a single-stage dense spatial and sparse temporal approach that considers complete spatial computation due to the uncertainty in the respiratory regions optimizing temporal computations by targeting

Conference acronym 'XX, June 24–26, 2025, Manhattan, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

motion rather than tracking long-range movement. Next, we develop a robust respiratory waveform tracking model tailored for sparse, independent, long-range point tracking, allowing for interpretable, adaptable design choices without requiring global motion adjustments [63] or test-time optimization [63]. After evaluating various TAP models, we identify deep learning particle tracking (PIPs)[22, 66] as a promising solution and repurpose the PIPs model to develop a specialized particle tracking model to achieve accurate respiratory waveform reconstruction, addressing key challenges such as large-scale data annotation, multi-modal data processing, multi-scale tracking, and precise localization of slight non-rigid pixel movements. To our knowledge, this research is the first to address respiratory waveform estimation using a TAP solution.

We pose our contributions as follows:

- **Coarse-to-fine scale RR point tracking**: We introduce a coarse-to-fine (top-down) continuous, high-fidelity respiratory waveform reconstruction across multiple video modalities. It accurately localizes dominant respiratory-induced movement regions (ROIs) through motion magnification, optical flow estimation, and coarse-scale tracking from a few high-resolution video frames and analyzes their respiratory signal strength. Secondly, we upscale the ROI locations and perform precise fine-scale respiratory movement tracking. This interpretable estimation addresses (a) the "where"—visually depicting their trajectories and (b) the "how" -during RR calculations from video frame sequences.
- **RRPIPs**: We develop a specialized particle tracking model, RRPIPs, and train them on multimodal respiratory videos to capture respiratory waveforms. This model can track single-pixel (coarse) and multi-pixel (fine) respiratory displacement continuously across frames, performing robustly across various video modalities (RGB, NIR, IR) and frame rates.
- **Dataset and experimentation**: We curated a large-scale, real-world respiratory pixel tracking video dataset by combining two public (Sleep, AIR) datasets with **in-house** respiratory videos, specifically for developing and validating video-based respiratory waveform models. Using an efficient, semi-supervised human-in-the-loop annotation process, we generate high-quality ground truth labels for non-rigid respiratory-induced displacements using optical flow and point-tracking models. The dataset includes diverse demographics, respiratory patterns, body regions, and environmental conditions. Our experiments achieve RR estimation accuracy within an MAE of 1, outperforming existing baselines, and extensive ablation studies further validate our approach. The dataset and code are open-sourced to support further RR research.

2. Background

Respiration Induced Movement in the Videos: The process of breathing involves rhythmic, continuous, and mostly involuntary movements of the diaphragm (a thin dome-shaped muscle positioned beneath the lungs and heart) and rib muscles, resulting in two phases: inspiration and expiration (as illustrated in Figure 1 (a)). During inspiration (breathing in or inhaling), the diaphragm contracts, descending, while rib muscles simultaneously contract, expanding the chest cavity. This coordinated effort increases lung volume, leading to a reduced lung air pressure that draws air into the lungs. In contrast, in the expiration (breathing out or exhaling) phase, the diaphragm relaxes, retaining its dome-like structure, and the rib muscles relax, which increases lung air pressure and causes the expulsion of air from the lungs.

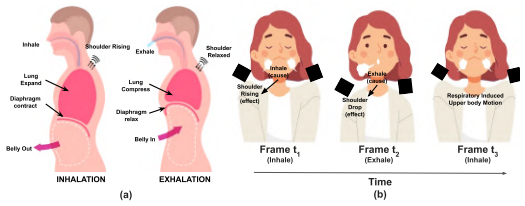

Figure 1: (a) Respiratory rate Mechanism during inhale and exhale phases. (b) Video cameras capture subtle respiration-induced movements in subsequent frames.

Problem Statement and Design Criteria: The objective is to accurately track respiratory-induced movements in body components and estimate the corresponding respiratory waveform from video data. This task is challenging due to the sparse and localized nature of respiratory motion, variance in signal strength between regions, motion expanding in few pixels, and low-texture surfaces of areas of the body involved in respiration (Figure 12). Our design criteria focus on developing a multimodal, multi-scale framework that leverages dense correlation maps for coarse-scale region identification and sparse tracking for fine-scale motion estimation. The model must efficiently handle uncertainty in the location of respiratory motion while ensuring high accuracy with minimal points in key regions. To achieve robust respiratory rate estimation, the model should support long-range, multi-frame tracking and be adaptable across diverse settings without requiring retraining or test-time optimization methods. It emphasizes flexibility, ease of retraining, and efficient computational design by avoiding dense operations in non-relevant video regions and focusing resources on accurately tracking respiratory areas.

Optical Flow using RAFT: To detect respiratory rate, it is essential to track the motion of external organs involved in respiration, such as the chest and abdomen. An effective solution is to compute optical flow, which estimates the motion between consecutive frames in a video by analyzing pixel displacements. Optical flow techniques enable the tracking of subtle, respiration-induced movements across frames. Among various optical flow methods, Recurrent All-Pairs Field Transforms (RAFT) [59] stand out for their precision and efficiency. RAFT computes dense optical flow by analyzing all-pair pixel correlations between two frames and iteratively refining the motion estimates through a recurrent neural network. It builds a high-resolution correlation volume to capture pixel interactions, enabling precise tracking even in low-texture or challenging conditions. RAFT's ability to handle fine-grained displacements makes it well-suited for respiratory motion tracking, where movements are often subtle and localized.

Video Motion Magnification: The motion of the outer organs caused by respiration is often very subtle, making it challenging to detect even with advanced optical flow models such as RAFT. Even if RAFT detects the optical flow, the magnitude of the motion is often too low to be effectively analyzed. Video motion magnification addresses this limitation by amplifying these small, imperceptible

motions within specific frequency ranges associated with respiration. Learning-based approaches, such as DeepMag [49], utilize deep convolutional neural networks (CNNs) to achieve this. The network is trained on synthetic data designed to capture small motions, with three main components: an encoder for spatial decomposition, a manipulator to magnify motion by scaling differences between frame representations, and a decoder to reconstruct the magnified frames. By enhancing subtle respiratory displacements, video motion magnification improves the detectability and accuracy of subsequent motion tracking under low signal-to-noise conditions.

Particle Video for video motion estimation: Video motion estimation aims to determine pixel movements across consecutive frames and can be broadly classified into six groups: sparse feature tracking, optical flow, feature matching, pixel-level long-range tracking, video-based motion optimization, and neural video representations [63]. Sparse feature tracking requires predefined features, which may not be available in all cases, while optical flow estimates a comprehensive motion field but can be computationally expensive for our framework. The tracking-all-points (TAP) problem generalizes motion estimation by tracking any given point across frames to reconstruct its trajectory. Recent deep learning frameworks, such as PIPs and PIPs++[22, 66], enhance TAP by integrating sparse feature tracking and dense optical flow, enabling persistent tracking through occlusions and leveraging multi-frame temporal priors for accurate long-range particle tracking. However, these models face challenges with small-scale oscillatory movements, less-textured surfaces, and domain shifts due to reliance on synthetic RGB data, limiting their generalization to diverse real-world scenarios[6, 31]. In the context of respiratory videos, we reinterpret respiration-induced organ movements as oscillating particles across frames and adapt PIPs++ for dense pixel-level tracking. Fine-tuning PIPs++ addresses its limitations and ensures applicability to real-world respiratory motion tasks, where precision demands incorporating local spatial context around the target pixels. Following [22], we use "point" and "particle" interchangeably, while "pixel" denotes discrete image grid cells.

3. Related Work

Existing RR estimation methods broadly fall into two main categories: signal processing and data-driven. Signal processing methods involve selecting different ROI, such as the chest or abdomen, through strategies such as manual cropping [46], Viola-Jones face detection algorithm [56, 61], cascade face classifier [27] and applying diverse techniques like filtering [5, 23, 27, 46], Eulerian video magnification [4, 8, 51], motion detection [34], intensity variation [37, 38, 41], spectral subtraction with canonical correlation analysis [9], marker tracking [2, 17, 36], auto-regressive models [58], and optical flow between frames [29, 42, 53].

Alternatively, recent research has shown that DL models driven by data have RR estimation, offering heuristic-free solutions. DL approaches, utilizing architectures such as Convolutional Neural Networks and signal decomposition [21, 28, 32], Artificial Hydrocarbon Network [8], flow-based network [35], single-shot object detection [33], YOLO object detector [43], and human pose detectors [18], showcase the potential to extract RR frequency from videos in an end-to-end fashion. In this research, in contrast to existing solutions, we propose a hybrid approach of DL-based point tracking and simple single processing for RR estimation from tracked point trajectories. To our knowledge, this is the first particle video algorithm adoption for video-based RR estimation.

The existing RGB-based RR dataset mostly covers infants' RR patterns during sleeping and mostly private datasets [35]. There are currently limited publicly available multipurpose resources in the adult video RR dataset, such as synthetic dataset SCAMPS [45], real datasets COHFACE [24], MAHNOB [57], NIR and thermal dataset [27] mostly covering the front shoulder movements of sitting subjects. Alternatively, we aim to develop and open-source our comprehensive video-based RR dataset specialized for RR estimation on adults covering multiple respiration-induced organs' movement with practical variations to facilitate general use cases and further research.

4. Methodology

We formulate video-based respiratory waveform estimation as localizing and tracking subtle, respiration-induced particle movements across frames. These oscillatory pixel displacements are confined to specific regions, while most of the frame remains static. Here, we propose a coarse-to-fine (top-down) framework decomposing the problem into two steps: (i) identifying respiratory motion regions by analyzing all spatial areas, as the locations of respiratory organs and their motions are unknown in uncontrolled videos, and (2) tracking fine-scale pixel movements within these regions to reconstruct the waveform. The framework operates in three stages, progressively narrowing spatial focus while expanding temporal analysis, ensuring efficient and accurate estimation as detailed below.

4.1. Respiratory Region Localization

This stage addresses the challenge of identifying the regions of interest (ROI) for respiratory motion in a video, where respiratory-induced movements can occur at any spatial location. We use a single-stage one-time dense spatial and sparse temporal operation to achieve efficient and accurate localization. Instead of tracking motion over time, we focus on localizing respiratory motion by analyzing two randomly selected frames taken a short time apart. This method minimizes temporal computations while ensuring comprehensive spatial analysis.

Motion Magnification: We apply motion magnification to amplify subtle respiratory-induced movements that might span only a few pixels. We use the pre-trained DeepMag model [49], which takes two frames as input and outputs a magnified motion frame. This magnification transforms small pixel displacements into more noticeable movements, facilitating effective region detection.

Optical flow estimation: We estimate dense optical flow between the original and magnified frames to detect pixel-wise motion. Here, we leverage the pre-trained RAFT model [59], to compute a dense optical flow field for each pixel comprising two channels representing x- and y-axis movements. We generate a 2D heatmap for the flow field by calculating the magnitude (the root of the squared sum of the x- and y-components). This heatmap highlights the spatial regions with the most significant movement as in Figure 3. We then apply a threshold (75th percentile) to select the regions with the highest flow magnitude, masking out irrelevant areas. The

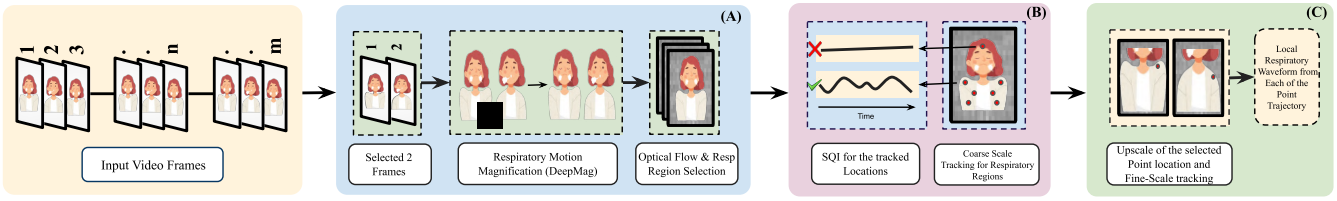

Figure 2: (a) Sparse temporal and full spatial operation (b) Medium temporal and Medium Spatial Operation (c) Full Temporal and Sparse spatial operation.

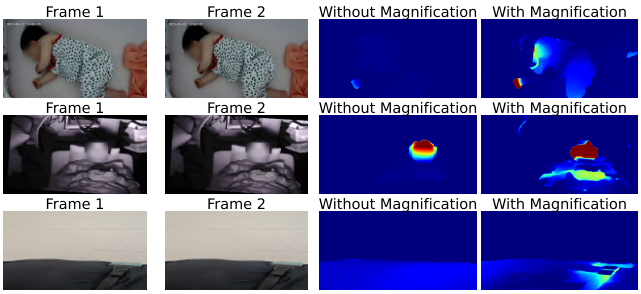

Figure 3: The columns represent 1st Frame, 2nd frame, optical flow magnitude between original frames, and between original and magnified frame. The magnification enhances the respiratory movements resulting in better RAFT performance.

resulting spatial regions represent the initial ROI containing potential respiratory-induced motion. These ROIs serve as the key spatial areas for subsequent stages of respiratory waveform estimation.

4.2. Quality Respiratory Motion Localization

In this stage, we refine the regions identified in Stage 1 by pinpointing specific coordinates with the strongest respiratory-induced motion. This involves intermediate spatial and temporal operations, combining coarse-scale point tracking with signal quality analysis to identify points with reliable respiratory movement.

Coarse-Scale point tracking: To improve resolution, the selected regions of the RAFT model are upscaled by a factor of two. Within these regions, multiple random points are selected and tracked over a limited number of frames using the developed RRPIPs model, which is specifically designed to handle subtle oscillatory movements caused by respiration as discussed in the following section. The trajectories obtained from this step serve as the basis for identifying locations of significant respiratory signals.

Signal Quality Index (SQI)-Based Coordinate Localization: Not all points within the region exhibit equal respiratory signal quality. We employ a Signal Quality Index (SQI) block that evaluates the SNR of the coarse-scale trajectory of the tracked points in terms of RR information to rank the points to select point coordinates with high-quality and reliable respiratory-induced motions for the final stages of fine-scale motion tracking and waveform reconstruction. Our SQI block consists of three steps of (i) *Waveform Extraction:* Each point's trajectory is converted into a respiratory waveform using the method discussed in later sections. (ii) *Spectral Analysis:* We compute the Power Spectral Density (PSD) of each point's waveform, focusing on the respiratory frequency range, and (iii) *Point Ranking:* Points are ranked based on the PSD magnitude and signal energy near the dominant respiratory frequency. This

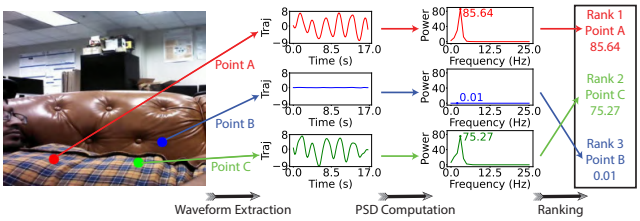

Figure 4: Signal Quality Index (SQI) Framework for Point Selection Based on Respiratory Signal Quality

coarse-level SQI estimation further narrows down potential areas of interest, focusing on points with the highest power in the relevant frequency band for subsequent fine-scale tracking.

4.3. Fine-Scale Point Tracking and Respiratory Waveform Extraction

In this stage, we refine the respiratory signal extraction process focusing on the dominant respiratory motion points identified in Stage 2. This involves upscaling the selected regions for fine-scale point tracking, ensuring accurate respiratory waveform reconstruction.

Region Upscaling: To facilitate fine-scale tracking, we crop a small window centered on each selected point from the dominant respiratory regions. These cropped regions are upscaled by a factor of eight to enhance resolution, magnifying previously subtle movements spanning only a few pixels. The spatial resolution is increased while maintaining the frame resolution, enabling more precise tracking of minute, non-rigid respiratory-induced pixel displacements.

Fine-Scale Point Tracking: We perform sparse spatial operations combined with full temporal operations over the upscaled regions. Using the same RRPIPs model as in Stage 2, we conduct multi-frame tracking to capture the fine-grain oscillatory pixel movements caused by respiratory organ motion. This continual fine-grain tracking addresses the challenges of subtle, non-rigid motion and ensures accurate trajectory estimation for each point.

Respiratory Waveform Extraction: We extract the respiratory waveforms from the fine-scale point trajectories for the upscaled regions. This top-down method, starting from coarse localization and progressing to fine-grain tracking, enhances both the accuracy and interpretability of the respiratory waveform extraction process, ensuring reliable and robust respiratory waveform estimation.

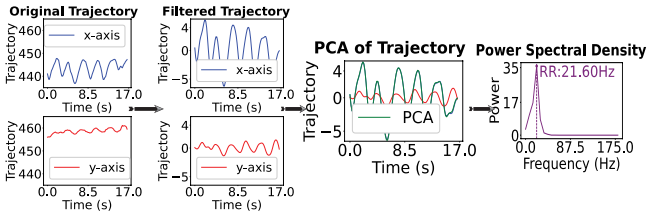

Figure 5: Steps to extract respiratory waveform from Point trajectory.

5. RRPIPS

This section introduces RRPIPS, a generalized multimodal, multi-scale persistent independent particle model for estimating respiratory waveforms by tracking respiratory-induced movements. Built on the PIPs++ framework, RRPIPS addresses the challenges of video-based respiratory waveform estimation through tailored enhancements and fine-tuning using curated real-world multimodal respiratory datasets.

The PIPs++ model [22, 66] offers a strong foundation due to its balance of flexibility, accuracy, computational efficiency, and suitability for respiratory motion tracking. Its CNN-based architecture supports sparse, long-range point tracking within small temporal windows, effectively capturing subtle, non-rigid oscillatory respiratory movements. The model's iterative refinement mechanism improves tracking accuracy, and its correlation maps provide interpretability, fostering trust and enabling domain experts to identify and address errors. However, its original training on synthetic RGB datasets limits its generalizability to real-world multimodal applications involving RGB, NIR, and IR data and subtle non-rigid respiratory motion.

RRPIPS extends PIPs++ with key adaptations. First, it incorporates multimodal generalization for robust tracking across diverse data sources. Second, it introduces multi-scale mechanisms to handle coarse global shifts and fine local displacements. Finally, fine-tuning on a curated multimodal, multi-scale respiratory dataset ensures accurate and reliable performance in real-world scenarios.

5.1. RRPIPS Development

We developed RRPIPS, a multi-frame tracking model specialized in respiratory-induced motion tracking across multiple scales and modalities. Built upon the PIPs++ architecture, the model processes 32 input frames per pass at varying spatial resolutions. To adapt PIPs++ for this task, we fine-tuned it using a curated multimodal real-world respiratory point-tracking dataset, ensuring robust and precise respiratory waveform estimation.

Training Data Preparation: The training dataset was semi-automatically created from multimodal respiratory motion data, incorporating RGB, NIR, and IR modalities. Ground truth was generated using a semi-supervised, human-in-the-loop optical flow-based labeling strategy, which ensured efficient and accurate data preparation, as discussed later. The training dataset preparation process follows the PIPs ++ training pipeline, in which a segment of 32 frames is selected and 128 points are placed in the initial frame, with their tracking information derived from our labeling strategy to enable supervised, end-to-end training. To enhance robustness, we augmented the dataset while preserving respiratory tracking information that includes rotation, frame shifting, rescaling, brightness, hue and saturation augmentation, multi-resolution training, color transformations, noise jittering, and salt-and-pepper noise.

Training Protocols and Strategies: The RRPIPS model was fine-tuned using multiple strategies to optimize its performance: (i) full training of the entire model from scratch, (ii) fine-tuning the entire pre-trained PIPs++ model, and (iii) component-specific tuning, which involved refining either the MLP mixer or the CNN feature encoder while freezing other components. In the latter approaches, smaller learning rates were employed to maintain training stability and prevent overfitting. The training protocol also included adjustments to emphasize respiratory-induced motion, enabling the model to generalize across multiple data modalities and scales. These strategies collectively enhance RRPIPS's capability for generalized tracking of respiratory movements and estimating respiratory waveforms from video data.

5.2. Tracking Operation

Data Preparation (Temporal Downsample): The RRPIPS model employs temporal downsampling to optimize computational efficiency while maintaining signal fidelity. Respiratory rate (RR) signals, typically ranging from 0.166 to 1 Hz, correspond to RR values across age groups (e.g. 12˘20 for adults over 18 years) [19, 54]. A minimum Nyquist frequency of 2 Hz is required to capture these signals without aliasing. Standard video cameras capturing at 30 fps exceed this threshold. Frames are selected at intervals of d ($d << 15$), producing a sequence with a sampling rate of $N_f = \frac{\text{fps}}{d}$, which remains above the Nyquist frequency. This approach reduces computational complexity by a factor of d while ensuring adequate performance.

Data Preparation (Spatial Upscale): Spatial domain processing enhances tracking accuracy through upscaling. For coarse-scale tracking, regions of interest (ROI) are extracted from RAFT model outputs and upsampled by a factor of two. For fine-scale tracking, localized regions around points selected using the Signal Quality Index (SQI) block are upsampled by a factor of eight. While both scales use the same tracking mechanism, coarse-scale processing uses masked outputs, and fine-scale processing uses windowed frames. These adjustments enable for accurate and efficient motion estimation across spatial resolutions.

Tracking Iteration: RRPIPS processes T frames per iteration to estimate pixel motion. For a downsampled video with N_f frames, the model requires $\frac{N_f}{T-1}$ iterations, as the last frame of each iteration serves as the first frame for the next (Figure 6).

In the j-th iteration, the model inputs T consecutive frames $[f_1, f_2, \ldots, f_T]^j$ and the coordinates of P pixels $(x_1^i, y_1^i)^j$ in the first frame, where $i \in \{1, 2, \ldots, P\}$. It tracks these P points across the remaining $T-1$ frames, outputting $(x_t^i, y_t^i)^j$ for $i \in \{1, 2, \ldots, P\}$ and $t \in \{2, \ldots, T\}$ (Figure 7).

During the first iteration ($j = 1$), P points are uniformly sampled across the selected pixel grid (Figure 2(a)). For subsequent iterations ($j > 1$), the tracked coordinates $(x_T^i, y_T^i)^j$ from the last frame of the j-th iteration initialize the starting coordinates $(x_1^i, y_1^i)^{j+1}$ for the $(j+1)$-th iteration. Similarly, the last frame f_T^j of the j-th iteration becomes the first frame f_1^{j+1} of the next iteration, ensuring seamless continuity in trajectory tracking across iterations.

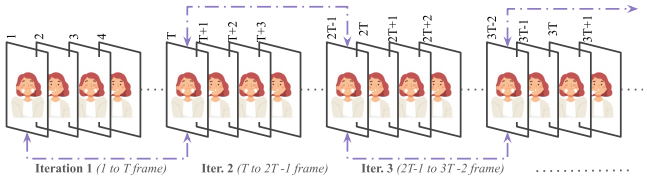

Figure 6: The model takes consecutive T frames to perform each iteration. The next iteration starts from the last frame of the previous iteration for tracking continuity.

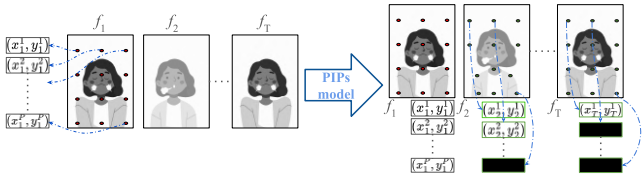

Figure 7: The PIPs model takes T frames and initial coordinates as input and estimates the pixel coordinates in the subsequent $T-1$ frames. The blue dotted lines represent the corresponding points' coordinates.

5.3. Respiratory Waveform from Point Trajectories:

Pixel Trajectory Estimation: For each point i, the RRPIPS model provides trajectory coordinates (x_j^i, y_j^i), where j represents the time-stamp of the point's location. To derive displacement over time, we calculate $(x_j^i - x_1^i)$ and $(y_j^i - y_1^i)$, resulting in two-channel time-series signals corresponding to movement in the x and y directions. Next, we extract movements on the respiratory frequency range, typically between 0.1 Hz and 1.5 Hz. A bandpass filter is applied with a passband frequency range of 0.05 Hz to 2 Hz, isolating the signal within the respiratory frequency band while removing noise and irrelevant low-frequency components. The filtered result is the estimated respiratory waveform, which provides the respiration-induced motion. We perform Power Spectral Density (PSD) analysis on the estimated respiratory waveform to determine the respiratory rate (RR). The maximum peak in the PSD indicates the frequency with the highest power within the respiratory range, corresponding to the RR. To extract the dominant motion component, we apply Principal Component Analysis (PCA) on these two filtered signals and select the first principal component, which captures the major variation in point trajectories (Figure 5).

6. Experiments

We perform a comparative analysis by experimenting with two existing approaches and naive baselines using our **In-house dataset**. We further experiment with two public datasets.

6.1. Dataset

This section describes the datasets used in our study.

6.1.1. In-house Dataset: In this study, we develop a video-based respiratory rate (RR) estimation dataset that mimics a clinical laboratory setting, capturing diverse test subjects, postures (sitting, standing, lying), RR patterns (regular, slow, fast), respiration-induced organ movements, and realistic environments. The dataset consists of 38 video sessions (3 to 5 minutes each) from 29 adult volunteers, recorded using a mounted stand camera and a static robot-mounted camera (ROSBot) positioned approximately 4 feet from the participants to capture RGB videos of shoulders, abdomen, and chest

Figure 8: Cropped sample frames from our in-house dataset

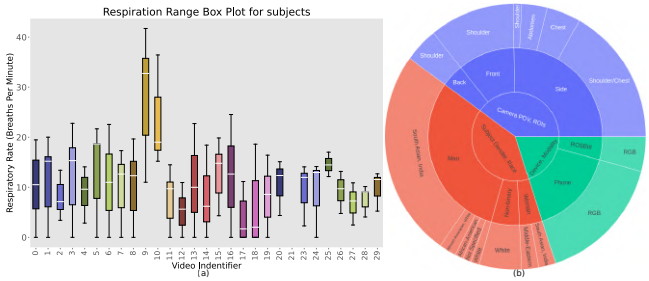

Figure 9: (a) Diversity in RR ranges (b) Diversity in Subjects and Camera POV and exposed RR induced organs.

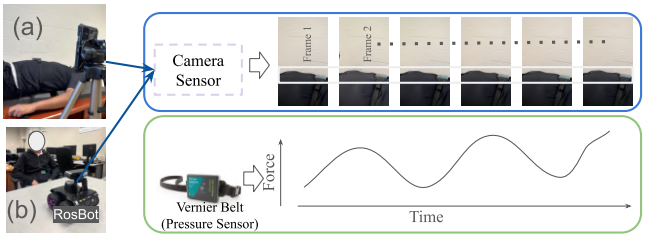

Figure 10: A sample data Collection Setting. The camera sensor captures the respiratory movements and the Vernier Belt provides ground truth respiration pressure. The imaginary white line acts as a visualization reference to highlight a subtle abdomen movement.

movements. Representative samples of the data are shown in Figure 8. Variations in participant demographics (age, gender, race), camera configurations (Mobile, ROSBot), backgrounds, noise levels, clothing, lighting, and tidiness enhance dataset diversity. A visual comparison of these variations is provided in Figure 9. Ground truth (GT) respiratory rate signals are measured using the Vernier Go Direct Respiration Belt [1], which tracks pressure changes during breathing. GT data is synchronized with video frames by marking the start of sensor data collection and verifying respiratory cycle correlations, with organ movement locations provided as metadata for each frame. An overview of the data collection setup is shown in Figure 10. Ethical considerations are also integral to our approach. We prioritize participant safety and privacy by obtaining informed consent, allowing them to review their data, and adhering to safety and privacy protocols. Additional guidelines ensure participant comfort and compliance during the data collection process, reinforcing our commitment to ethical research practices.

6.1.2. Public Dataset AIR-125 [35]: The AIR-125 dataset comprises 125 videos, each approximately 60 seconds long, featuring infants sourced from various real-world monitoring scenarios. RR annotations, ranging from 18 to 42 RR, were obtained from thoracic or abdominal motion. The dataset includes footage collected from

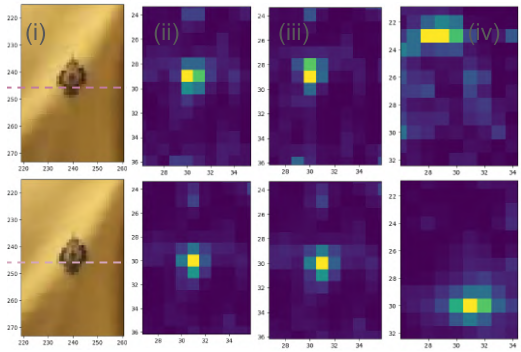

Figure 11: (i) column represents original frames. The red line shows a bit of pixel shifts. We observe the point shift in the feature spaces in the coarse scale in (ii) PIPS and (iii) RRPIPS. We also observe a visible shift in feature space for upscale data by RRPIPS.

baby monitors and YouTube, providing diverse resolutions and frame rates. **Sleep Databaset** [27]: Additionally, we utilized the Sleep Database, consisting of 28 videos featuring 11 adult subjects captured in dual-mode (NIR and IR) under illumination intensities ranging from 0 to 3 Lux.

6.2. Creating respiratory point Tracking Dataset

Annotating multi-modal-multiscale large datasets for respiratory waveform reconstruction poses considerable hurdles in attaining precision and scalability, particularly in practical applications. To address these challenges, we develop a semi-supervised annotation process that integrates many public and proprietary datasets. Leveraging advanced optical flow models and particle tracking algorithms, our approach automates the development of pixel displacement ground truth with minimal manual effort. We utilize the RAFT model, renowned for its robustness to large displacements, to capture coarse optical flow across frames and facilitate detailed respiratory motion annotations. Our method, inspired by co-tracker methodologies, bridges the gap between manual annotation and full automation, ensuring accurate analysis of real-world videos.

The annotation process begins with the selection of 32 frames from each video. Annotators then perform region of interest (ROI) selection, pinpointing areas of notable respiratory activity using grid overlays. Trajectory points are identified by designating grid centers, and in each ROI, we sample 128 points, with more than 80% concentrated in areas exhibiting significant motion. This approach enhances annotation accuracy while reducing computational burden. To efficiently handle dense optical flow processing, videos are segmented into smaller, manageable samples, ensuring scalability and streamlined analysis.

Annotation verification is supported by three intuitive visualizations: trajectory overlays on image plots, motion visualizations within ROIs, and trajectory start/end position plots. These provide instantaneous feedback, ensuring trajectory accuracy and minimizing errors. For fine-scale tracking, we employ the PIPs++ model, optimized on our multimodal, multi-scale dataset, to achieve pixel-level accuracy for non-rigid respiratory motion.

Overall, the datasets include a total number of 40,000 samples, each containing 32 frames, with each frame annotated with 128 trajectories. To ensure robust model evaluation, we employ a leave-subject-out testing strategy, splitting the dataset into training and testing sets to avoid subject overlap during evaluation. We optimize training with a multi-scale approach, leveraging data at different spatial resolutions to improve the model's adaptability to respiratory motion across scales. To further enhance generalizability, we employ extensive data augmentation techniques, including spatial transformations (shifting, axis transpositions, flipping), zoom-in effects, and pixel-level noise injection.

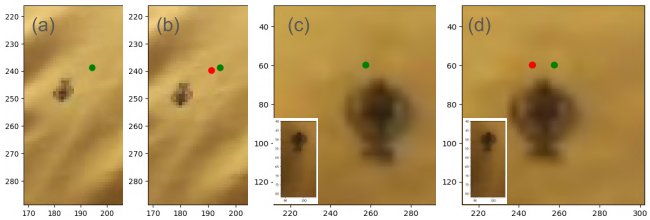

Figure 12: The (a) and (b) show movements in the original frame scale and (c) and (d) show pixel motion in the upscaled version. RRPIPS successfully tracks the points across scales.

6.3. Implementation Details

We conducted all experiments in a Python environment on a server with Intel(R) Core(TM) i9 processors, 128GB of RAM, and NVIDIA GeForce RTX 3090 GPUs.

Video Preprocessing: We used the OpenCV library for loading and preprocessing video files. Temporal downsampling was applied with factors of 2, 5, and 6, while spatial preprocessing involved cropping the region of interest (ROI) and applying a 2x upscaling factor during coarse-scale tracking. We crop a window of 96×128 centered around the selected points and upscale it to 384×512 spatial resolution for fine-scale tracking.

Magnification: We utilized the pre-trained DeepMag model for motion magnification in its static setting. The amplification factor was set to 10, which performed well across a range of 5 to 15.

RAFT Model: Optical flow was estimated using the RAFT-large model with default pre-trained weights. RAFT performed 12 iterations to compute the flow between frames. The input frames were resized to 520×960 and normalized to a range of $[-1, 1]$. To ensure robust results, we magnified and calculated optical flow for multiple sets of randomly selected two frames from early videos and retained regions consistently selected across iterations for tracking that avoids relying on selecting two erroneous frames in videos where respiratory-induced movement is subtle or distributed.

Pre-trained Models and Data: We leveraged pre-trained architectures of PIPs++, trained on the FlyingThings and Point Odyssey datasets, for fine-tuning the respiratory dataset. The model tracked 32 consecutive frames of size $380 \times 640 \times 3$ per iteration. Query points (32, 64, 128) were initialized in the ROI for coarse-scale tracking (refer to Figure 15).

Trajectory Analysis: We designed a Butterworth bandpass filter with a frequency range of 0.05 Hz to 2 Hz using the SciPy library to process point trajectories. Power Spectral Density (PSD) estimation was performed using Welch's method [64], and the top 6 points with the highest PSD peaks were selected for fine-scale tracking.

6.4. Ablation Studies

We conducted comprehensive ablation studies to evaluate the critical components of our pipeline and investigate alternative design choices. First, we assessed the significance of each stage in the proposed coarse-to-fine framework, including motion magnification, RAFT-based optical flow for region selection, the SQI block, upscaling, and model fine-tuning, demonstrating their individual and collective contributions to accurate respiratory waveform estimation. Second, we explored alternative design implementations to assess their impact on performance, such as replacing RAFT with OpenCV optical flow for motion tracking, utilizing phase-based magnification instead of DeepMag, substituting PIPS++ with the PIPS model, and testing configurations without fine-tuning. These experiments revealed valuable insights into the trade-offs between computational efficiency and accuracy. Finally, we optimized the RRPIPS model by examining the role of multimodal, multiscale data, model architecture, and various training strategies. We compared performance across various fine-tuning and optimization approaches, highlighting the advantages of tailoring the model to the respiratory dataset for enhanced accuracy and robustness.

6.5. Baselines

We implemented several methods for respiratory rate (RR) estimation, including intensity-based and optical-flow-based techniques [37, 41]. These methods leverage variations in pixel intensity and displacement, respectively, to capture subtle respiratory-induced movements. Additionally, we explored edge-based displacement tracking with manual region-of-interest (ROI) selection to improve accuracy.

Intensity-Based RR Estimation [37]: This method involves manually selecting an ROI and tracking pixel intensity gradients over time. The intensity variations are filtered, and frequency domain analysis is performed to estimate RR from the temporal changes in pixel intensity. **Optical-Flow-Based RR Estimation** [41]: Optical flow vectors are computed between successive frames using OpenCV's flow estimator. RR is extracted by reconstructing the flow over the entire video sequence. We conducted two experiments: one using raw frames and another using edge-detected frames to enhance displacement tracking. **Edge Shift Tracking:** We manually cropped ROI edges and projected them onto the X and Y axes. Respiratory-induced movements were captured by analyzing coordinate displacements of the edge projections across frames. **Area Under Edge Curve:** A small, continuous edge block was manually cropped, and the area under the edge curve was calculated across successive frames. The variation in this area was analyzed to estimate RR, capturing respiratory movements indirectly through spatial changes.

7. Results

7.1. Main Results

We evaluated our method on the left-out test dataset comprised of RGB, NIR, and IR video modalities from different datasets. Results were analyzed in both the time and frequency domains to assess tracking and respiratory rate (RR) estimation performance.

Time-Series Results: To evaluate tracking performance, we compared the point trajectories generated by our RRPIP model

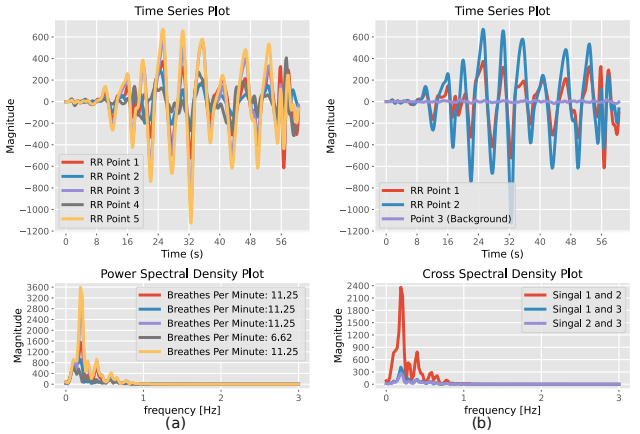

Figure 13: (a) Time (top) and frequency (bottom) domain visualization for selected ROI points (b) The high correlation between RR ROI pixels in time (top) and frequency domain (bottom).

with those from the RAFT optical flow model. This comparison highlights the ability of our model to maintain accurate tracking over multi-frame intervals, particularly in challenging scenarios involving non-rigid, subtle respiratory motions.

Frequency Domain Results: For RR estimation, we performed a quantitative analysis using standard metrics, including mean absolute error (MAE) and root mean square error (RMSE), reported as breaths per minute. For our **in-house** dataset, we followed the Vernier machine configuration, calculating RR over 30-second intervals with a 20-second overlap between stages. The final performance metrics were averaged across all subjects, providing a comprehensive assessment over the complete dataset.

Performance on In-House Dataset: Our approach outperformed baseline methods in estimating RR from the **in-house** dataset (Table 1). While intensity-based and edge-based solutions provided competitive RR estimates, they relied heavily on external blocks for ROI selection and edge detection. For edge-based methods, we optimized OpenCV's Canny edge detection parameters to achieve optimal results. In contrast, our model seamlessly integrates these functionalities, eliminating dependencies on manual parameter tuning. Additionally, two-frame optical flow (OF) methods were explored for pixel displacement detection, but were less robust in comparison to our coarse-to-fine pipeline.

Generalization to Public Datasets: We further assessed the adaptability of our approach on two public datasets featuring RGB videos of infants and NIR/IR videos of adults. Tables 4 and 3 present the performance metrics on these datasets. RRPIP achieved high accuracy in trajectory estimation for both NIR and IR modalities, demonstrating consistent RR estimation accuracy across test samples from diverse conditions.

Infant and Sleep Dataset Analysis: Our model performed competitively on RGB videos of infants. Notably, in the sleep dataset, RRPIP showed resilience to head movements, which typically involve small spatial areas within the frame. By leveraging multiple points for RR estimation, our method mitigates errors from individual point tracking failures, ensuring robust performance under such conditions.

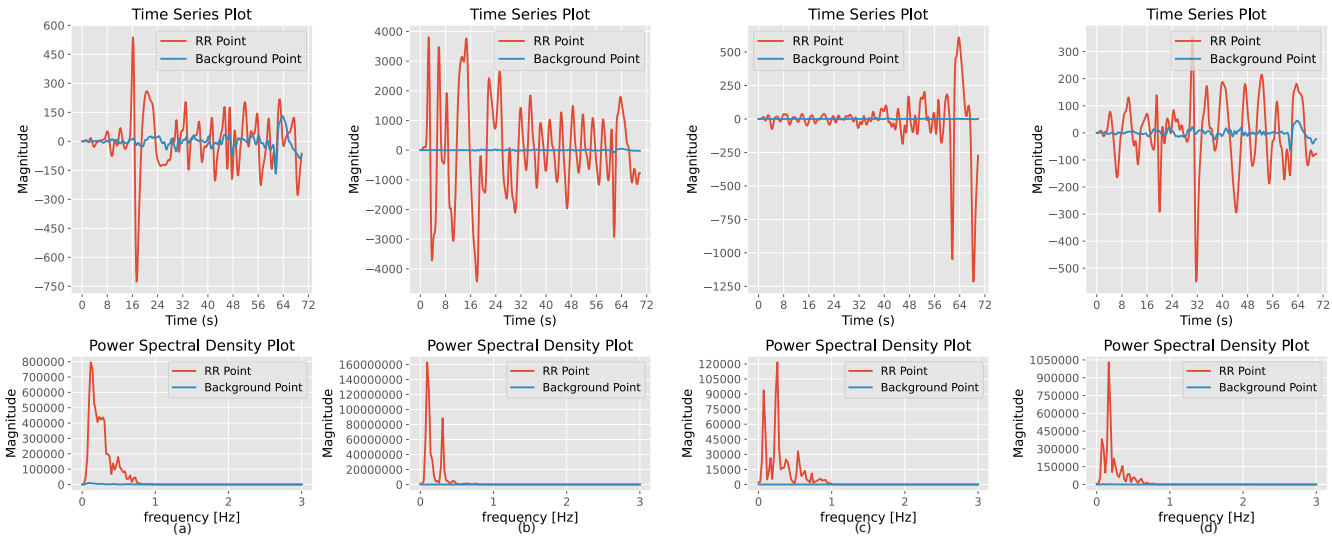

Figure 14: The top row shows the time domain plot of filtered trajectories and the bottom row shows the corresponding PSD. We observe the RR corresponding points result in higher energy in the time domain and ROI frequency region.

Table 1: Comparative results of RR estimation systems on our in-house RGB adult dataset.

Metrics	Intensity	OF	Edge Shift	Edge Area	pips++	RRPIPS
MAE	3.25	2.42	2.52	3.45	1.62	1.01
RMSE	5.12	3.89	5.53	6.82	2.92	1.80

Table 2: Summarized results on Public datasets.

Metrics	Sleep Data (Adult NIR)	Sleep Dataset (Adult IR)	AIR-125 (Infant RGB)	In-House (Adult RGB)
MAE	2.56	1.02	1.36	1.01
RMSE	4.70	2.12	2.75	1.80

Table 3: Comparative results RMSE of RR with different approaches. FT: full model training, HT: MLP layer training

Dataset	PIPS	PIPS++	RRPIPS(FT)	RRPIPS (HT)
Sleep-NIR	9.8	7.43	2.56	3.44
Sleep-IR	4.8	2.3	1.02	2.05
AIR	3.82	2.4	1.36	1.9
In-house	1.62	1.88	1	1.1

Table 4: Comparative results of RMSE with RAFT estimated target points with different approaches.

Dataset	PIPS	PIPS++	RRPIPS(FT)	RRPIPS(HT)	RRPIPS(RGB)
Sleep-NIR	12.8	9.2	3.5	4.2	8.7
Sleep-IR	10.1	6.8	2.1	2.7	6.9
AIR	7.2	4.2	2.7	3.5	3.8
In-house	7.8	3.5	3.6	2.5	2.4

The experimental results highlight the efficacy of our RRPIP framework across modalities and datasets, providing reliable and interpretable RR estimation and respiratory waveform extraction. Comprehensive quantitative metrics underscore the advantages of our multi-frame, coarse-to-fine pipeline in addressing the challenges of contactless respiration monitoring.

We also analyzed the qualitative performance and interpretability of the proposed RRPIPS framework by examining the spatial location and behavior of selected points as regions of interest (ROI). Our empirical observations of the selected points on source video

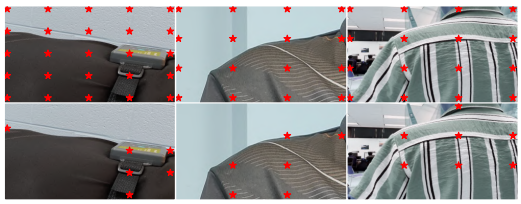

Figure 15: Top row shows the position of initial points and the bottom row shows the position of top-6 PSD ranked points.

frames (Figure 15) revealed that most were chosen from respiratory ROI, validating our design choices. This ensures that the respiratory rate (RR) estimation originates primarily from trajectories associated with external respiratory movements, enhancing interpretability and reliability.

Additionally, we examined the temporal and frequency characteristics of the selected points' trajectories. As depicted in Figure 16, the trajectories exhibited high correlation, even when points were selected from different locations or organs. These correlated trajectories aligned closely with respiratory movements in both the time and frequency domains, further demonstrating the framework's ability to capture meaningful respiratory signals.

To evaluate the multiscale tracking capability of RRPIPS, we investigated its performance across various spatial scales, from tracking motion between only a few pixels to capturing larger-scale displacements (Figure 12). The results underscore the robustness of the model in handling subtle, non-rigid respiratory motions. Finally, we visualized the tracking points at coarse and fine scales to provide a deeper understanding of the pipeline's behavior. Representative examples are shown in Figures 17 and 18, illustrating the transition from initial coarse-scale tracking to fine-scale tracking for precise respiratory waveform estimation. These visualizations confirm the effectiveness of RRPIPS in leveraging multi-scale information to achieve accurate and interpretable RR estimation.

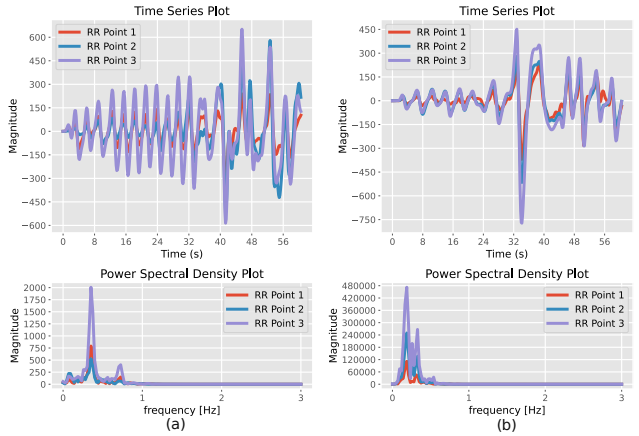

Figure 16: Movement pattern similarity in time and frequency domain for points over the RR pixels.

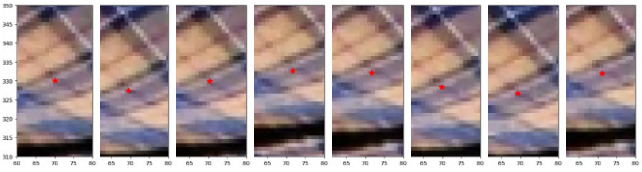

Figure 17: sample tracking in original video resolutions. The red point shows the selected point.

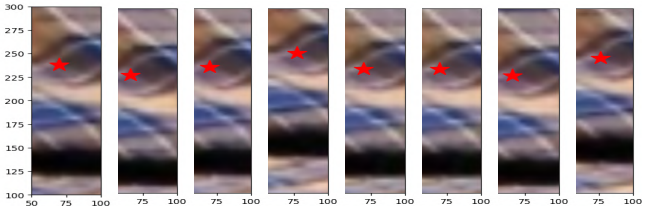

Figure 18: sample tracking in upscaled video resolutions in the selected regions

7.2. Ablation Studies

We conducted comprehensive ablation studies to evaluate the contributions of individual components and alternative design choices in our proposed pipeline, as illustrated in Figure 2.

Component Analysis: The region selection process, selecting the top 70th percentile of pixels with the highest respiratory activity, effectively reduces computational load by narrowing the tracking scope. By analyzing only two frames at extreme minima, this approach minimizes the need for exhaustive computations over long durations, relying instead on a one-time application of the DeepMag and RAFT optical flow models. The Signal Quality Index (SQI) block further refines this selection by isolating high-SNR points within respiratory regions, reducing the number of points tracked, and localizing regions for fine-scale upscaling. The upscaling operation enhances resolution, significantly improving trajectory estimation accuracy, as illustrated in Figure 11, particularly in distinguishing features for precise respiratory waveform extraction.

Role of Motion Magnification: Motion magnification strengthens RAFT-based optical flow estimation under challenging conditions, such as low motion scenarios or small frame gaps. While RAFT performs sufficiently well when visible respiratory motion exists between distant frames, magnification ensures consistency across diverse samples. However, openCV optical flow ROI selection performance degrades without motion magnification. Future studies will investigate adaptive criteria for applying magnification based on motion intensity and frame conditions.

Alternative Design Choices: We explored alternative methods to optimize the pipeline. Substituting RAFT with OpenCV optical flow yielded comparable results when combined with motion magnification. However, RAFT demonstrated superior accuracy in regions with subtle respiratory motion due to its robust area selection capabilities. For tracking, we compared off-the-shelf PIPS models with RRPIPS. While performance on RGB datasets and upscaled frames was comparable, RRPIPS significantly outperformed PIPS in NIR and IR modalities, particularly for tracking small-scale respiratory particle movements (Table 4 and Table 3). The PIPS model underperformed in these modalities due to its general-purpose design for RGB videos, with NIR frames showing better outcomes than IR. Analysis of the dataset revealed that the IR frames, dominated by body heat radiation, and low-visibility NIR frames contributed to the observed discrepancies. We also evaluated phase-based video motion magnification [62] as an alternative to DeepMag. While phase-based magnification achieved comparable results when tuned to the respiratory frequency range, it required significantly more temporal video frames than the two-frame approach of DeepMag, increasing computational demands.

Data and Training Strategies: We investigated the impact of various fine-tuning strategies. Full fine-tuning of the model, including correlation cost map features, achieved the best results in both feature space representation and estimation accuracy (Table 3 and Table 4). Additionally, fine-tuning using multimodal datasets enhanced RRPIPS's performance in IR and NIR modalities, consistently outperforming the baseline PIPS model. We observe suboptimal performance in other modalities by training with only RGB modalities as in table 4 highlighting the importance of multimodal optimization in robust respiratory waveform tracking.

8. Discussion

Avoidance of Object Detection and Segmentation: Our framework eliminates the need for manual cropping, physical markers, or object detection. Object detection underperforms due to the lack of clear reference points, while segmentation models fail in scenarios where respiratory organs are covered or only motion is visible. For instance, our method successfully tracked respiratory-induced belly movements by covering hands or clothing.

Manual Points Selection: Our framework incorporates a human-in-the-loop approach for manual point selection to track specific respiratory organs. Experts can define regions of interest, allowing the model to focus on motion tracking in those areas. This enhances the detection of respiratory discomfort and the monitoring of expected respiratory movements with greater precision. In addition, the framework supports the selection of multiple points for tracking, enabling comparative analysis of respiratory waveforms across various respiratory-induced organs, making it highly versatile for clinical and research applications.

Efficiency of Sparse Tracking with PIPS: Using RAFT for dense tracking in respiratory waveform estimation is computationally expensive due to its 4-D correlation matrix calculations across

complete spatial regions and its two-frame operation per iteration. We employed RAFT only for one-time dataset generation, as it is effective for this purpose. In contrast, PIPS excels in sparse point tracking by utilizing local features, offering a more efficient solution without sacrificing accuracy in estimating respiratory motion.

Impact of Downsampling: The downsampling operation offers two advantages: enhanced computational efficiency and improved signal-to-noise ratio (SNR) for respiratory motion. By processing reduced frames, the model avoids high-frequency noise, retaining the slow-moving respiratory-induced signals that are otherwise drowned out in high frame-rate videos. This counterintuitive finding underscores the importance of adapting frame rates for optimal respiratory rate (RR) estimation.

Movement Constraints and Model Limitations: The framework assumes that respiratory-induced movements dominate within the frequency range of interest (0.1–1.5 Hz) and that background motion has minimal impact. These assumptions hold in controlled static subject-camera settings but may degrade performance with significant non-respiratory movements or subpixel-level motion. While filtering constraints ensure robustness against out-of-band motion frequencies, future work will address dynamic scenarios such as subject or camera motion, speech-induced artifacts, exercise, and regular activities. Additionally, we aim to establish motion limits for all models and improve performance in challenging cases by separating respiratory signals from other movements.

Future Directions: Extending the framework includes fine-tuning RAFT and DeepMag for non-RGB modalities, addressing model limitations in edge devices, and enhancing robustness to background interference. Incorporating global motion analysis (GMA) for separating respiratory and non-respiratory movements and enabling deployment in unconstrained environments will further broaden its applicability. Further, we will explore other test time optimization-based TAP models for volumetric representation and motion tracking [63].

9. Conclusion

This work addresses the challenge of estimating respiratory waveforms from videos capturing respiratory-induced motions for contactless health monitoring. We propose a novel coarse-to-fine pipeline that identifies respiratory motion regions using motion magnification and optical flow, localizes respiratory-induced pixel movements via coarse-scale tracking and a Signal Quality Index (SQI) block, and refines tracking through upscaling for precise waveform estimation. Our specialized RRPIPS model is designed for robust tracking across multiple modalities (RGB, NIR, IR) and scales, capturing subtle respiratory motions effectively. We also introduce a large-scale RGB dataset featuring diverse respiratory-induced movements in realistic settings, complemented by a semi-supervised, human-in-the-loop approach for generating ground truth trajectory annotations. Validated across three diverse datasets and modalities, our approach achieves state-of-the-art results, with extensive studies demonstrating its efficiency, adaptability, and potential for advancing contactless respiratory monitoring systems.

References

[1] [n. d.]. Vernier Go Direct Respiration Belt. https://www.vernier.com/ product/go-direct-respiration-belt/.

[2] Mona Alnaggar, Ali I Siam, Mohamed Handosa, T Medhat, and MZ Rashad. 2023. Video-based real-time monitoring for heart rate and respiration rate. *Expert Systems with Applications* 225 (2023), 120135.

[3] Christoph Hoog Antink, Simon Lyra, Michael Paul, Xinchi Yu, and Steffen Leonhardt. 2019. A broader look: Camera-based vital sign estimation across the spectrum. *Yearbook of medical informatics* 28, 01 (2019), 102–114.

[4] Luca Antognoli, Paolo Marchionni, Stefano Nobile, Virginio Paolo Carnielli, and Lorenzo Scalise. 2018. Assessment of cardio-respiratory rates by non-invasive measurement methods in hospitalized preterm neonates. In *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 1–5.

[5] Marek Bartula, Timo Tigges, and Jens Muehlsteff. 2013. Camera-based system for contactless monitoring of respiration. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2672–2675.

[6] Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yitong Dong, Yijin Li, and Hongsheng Li. 2023. Context-PIPs: persistent independent particles demands spatial context features. *arXiv preprint arXiv:2306.02000* (2023).

[7] Sheldon R Braun. 1990. Respiratory rate and pattern. *Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition* (1990).

[8] Jorge Brieva, Hiram Ponce, and Ernesto Moya-Albor. 2023. Non-Contact Breathing Rate Estimation Using Machine Learning with an Optimized Architecture. *Mathematics* 11, 3 (2023), 645.

[9] Juan Cheng, Runqing Liu, Jiajie Li, Rencheng Song, Yu Liu, and Xun Chen. 2023. Motion-Robust Respiratory Rate Estimation from Camera Videos via Fusing Pixel Movement and Pixel Intensity Information. *IEEE Transactions on Instrumentation and Measurement* (2023).

[10] Youngjun Cho, Simon J Julier, Nicolai Marquardt, and Nadia Bianchi-Berthouze. 2017. Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging. *Biomedical optics express* 8, 10 (2017), 4480–4503.

[11] C Chourpiliadis and A Bhardwaj. 2022. Physiology, respiratory rate.[Updated 2022 Sep 12]. *StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing* (2022).

[12] Matthew M Churpek, Trevor C Yuen, Seo Young Park, David O Meltzer, Jesse B Hall, and Dana P Edelson. 2012. Derivation of a cardiac arrest prediction model using ward vital signs. *Critical care medicine* 40, 7 (2012), 2102.

[13] Michelle A Cretikos, Rinaldo Bellomo, Ken Hillman, Jack Chen, Simon Finfer, and Arthas Flabouris. 2008. Respiratory rate: the neglected vital sign. *Medical Journal of Australia* 188, 11 (2008), 657–659.

[14] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. 2022. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems* 35 (2022), 13610–13626.

[15] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. 2023. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10061–10072.

[16] Heather E Elphick, Abdulkadir Hamidu Alkali, Ruth K Kingshott, Derek Burke, and Reza Saatchi. 2019. Exploratory study to evaluate respiratory rate using a thermal imaging camera. *Respiration* 97, 3 (2019), 205–212.

[17] Shamel Fahmi, Frank FJ Simonis, and Momen Abayazid. 2018. Respiratory motion estimation of the liver with abdominal motion as a surrogate. *The International Journal of Medical Robotics and Computer Assisted Surgery* 14, 6 (2018), e1940.

[18] Isaac René Aguilar Figueroa, Jesús Vladimir Martínez Nuño, and Eduardo Gerardo Mendizabal-Ruiz. 2019. Remote Optical Estimation of Respiratory Rate Based on a Deep Learning Human Pose Detector. In *Latin American Conference on Biomedical Engineering*. Springer, 234–241.

[19] Susannah Fleming, Matthew Thompson, Richard Stevens, Carl Heneghan, Annette Plüddemann, Ian Maconochie, Lionel Tarassenko, and David Mant. 2011. 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 (2011), 1011–1018.

[20] Michele Girardi, Andrea Nicolò, Ilenia Bazzucchi, Francesco Felici, and Massimo Sacchetti. 2021. The effect of pedalling cadence on respiratory frequency: Passive vs. active exercise of different intensities. *European Journal of Applied Physiology* 121 (2021), 583–596.

[21] Tianqi Guo, Qian Lin, and Jan Allebach. 2021. Remote estimation of respiration rate by optical flow using convolutional neural networks. *Electronic Imaging* 2021, 8 (2021), 267–1.

[22] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. 2022. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*. Springer, 59–75.

[23] MA Hassan, AS Malik, D Fofi, N Saad, and F Meriaudeau. 2017. Novel health monitoring method using an RGB camera. *Biomedical optics express* 8, 11 (2017), 4838–4854.

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

[25] Nadine Hochhausen, Carina Barbosa Pereira, Steffen Leonhardt, Rolf Rossaint, and Michael Czaplik. 2018. Estimating respiratory rate in post-anesthesia care

unit patients using infrared thermography: an observational study. *Sensors* 18, 5 (2018), 1618.

[26] Timothy J Hodgetts, Gary Kenward, Ioannis G Vlachonikolis, Susan Payne, and Nicolas Castle. 2002. The identification of risk factors for cardiac arrest and formulation of activation criteria to alert a medical emergency team. *Resuscitation* 54, 2 (2002), 125–131.

[27] Menghan Hu, Guangtao Zhai, Duo Li, Yezhao Fan, Huiyu Duan, Wenhan Zhu, and Xiaokang Yang. 2018. Combination of near-infrared and thermal imaging techniques for the remote and simultaneous measurements of breathing and heart rates under sleep situation. *PloS one* 13, 1 (2018), e0190466.

[28] Hyeonsang Hwang, Kunyoung Lee, and Eui Chul Lee. 2022. A real-time remote respiration measurement method with improved robustness based on a CNN model. *Applied Sciences* 12, 22 (2022), 11603.

[29] Rik Janssen, Wenjin Wang, Andreia Moço, and Gerard De Haan. 2015. Video-based respiration monitoring with automatic region of interest detection. *Physiological measurement* 37, 1 (2015), 100.

[30] Joao Jorge, Mauricio Villarroel, Sitthichok Chaichulee, Alessandro Guazzi, Sara Davis, Gabrielle Green, Kenny McCormick, and Lionel Tarassenko. 2017. Non-contact monitoring of respiration in the neonatal intensive care unit. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 286–293.

[31] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. 2024. CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos. *arXiv preprint arXiv:2410.11831* (2024).

[32] Fatema-Tuz-Zohra Khanam, Asanka G Perera, Ali Al-Naji, Kim Gibson, and Javaan Chahl. 2021. Non-contact automatic vital signs monitoring of infants in a neonatal intensive care unit based on neural networks. *Journal of Imaging* 7, 8 (2021), 122.

[33] Daniel G Kyrollos, Joshua B Tanner, Kim Greenwood, JoAnn Harrold, and James R Green. 2021. Noncontact neonatal respiration rate estimation using machine vision. In *2021 IEEE SAS*. IEEE, 1–6.

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

[35] Sai Kumar Reddy Manne, Shaotong Zhu, Sarah Ostadabbas, and Michael Wan. 2023. Automatic Infant Respiration Estimation from Video: A Deep Flow-Based Algorithm and a Novel Public Benchmark. In *International Workshop on Preterm, Perinatal and Paediatric Image Analysis*. Springer, 111–120.

[36] Carlo Massaroni, Eugenio Cassetta, and Sergio Silvestri. 2017. A novel method to compute breathing volumes via motion capture systems: design and experimental trials. *Journal of applied biomechanics* 33, 5 (2017), 361–365.

[37] Carlo Massaroni, Daniela Lo Presti, Domenico Formica, Sergio Silvestri, and Emiliano Schena. 2019. Non-contact monitoring of breathing pattern and respiratory rate via RGB signal measurement. *Sensors* 19, 12 (2019), 2758.

[38] Carlo Massaroni, Daniel Simões Lopes, Daniela Lo Presti, Emiliano Schena, and Sergio Silvestri. 2018. Contactless monitoring of breathing patterns and respiratory rate at the pit of the neck: A single camera approach. *Journal of Sensors* 2018 (2018).

[39] Carlo Massaroni, Andrea Nicolò, Daniela Lo Presti, Massimo Sacchetti, Sergio Silvestri, and Emiliano Schena. 2019. Contact-based methods for measuring respiratory rate. *Sensors* 19, 4 (2019), 908.

[40] Carlo Massaroni, Andrea Nicolo, Massimo Sacchetti, and Emiliano Schena. 2020. Contactless methods for measuring respiratory rate: A review. *IEEE Sensors Journal* 21, 11 (2020), 12821–12839.

[41] Carlo Massaroni, Emiliano Schena, Sergio Silvestri, and Soumyajyoti Maji. 2019. Comparison of two methods for estimating respiratory waveforms from videos without contact. In *2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 1–6.

[42] Marc Mateu-Mateus, Federico Guede-Fernández, Víctor Ferrer-Mileo, Miguel A García-González, Juan Ramos-Castro, and Mireya Fernández-Chimeno. 2019. Comparison of video-based methods for respiration rhythm measurement. *Biomedical Signal Processing and Control* 51 (2019), 138–147.

[43] Lalit Maurya, Reyer Zwiggelaar, Deepak Chawla, and Prasant Mahapatra. 2023. Non-contact respiratory rate monitoring using thermal and visible imaging: a pilot study on neonates. *Journal of Clinical Monitoring and Computing* 37, 3 (2023), 815–828.

[44] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on CVPR*. 4040–4048.

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

[46] Yunyoung Nam, Youngsun Kong, Bersain Reyes, Natasa Reljin, and Ki H Chon. 2016. Monitoring of heart and breathing rates using dual cameras on a smartphone. *PloS one* 11, 3 (2016), e0151013.

[47] Andrea Nicolò, Samuele M Marcora, and Massimo Sacchetti. 2016. Respiratory frequency is strongly associated with perceived exertion during time trials of different duration. *Journal of sports sciences* 34, 13 (2016), 1199–1206.

[48] Andrea Nicolò, Carlo Massaroni, Emiliano Schena, and Massimo Sacchetti. 2020. The importance of respiratory rate monitoring: From healthcare to sport and exercise. *Sensors* 20, 21 (2020), 6396.

[49] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik. 2018. Learning-based video motion magnification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 633–648.

[50] RS Nannan Panday, TC Minderhoud, N Alam, and PWB Nanayakkara. 2017. Prognostic value of early warning scores in the emergency department (ED) and acute medical unit (AMU): a narrative review. *European journal of internal medicine* 45 (2017), 20–31.

[51] Matthew Pediaditis, Cristina Farmaki, Sophia Schiza, Nikolaos Tzanakis, Emmanouil Galanakis, and Vangelis Sakkalis. 2022. Contactless respiratory rate estimation from video in a real-life clinical environment using eulerian magnification and 3D CNNs. In *2022 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE, 1–6.

[52] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 724–732.

[53] Leonardo Queiroz, Helder Oliveira, Svetlana Yanushkevich, and Reed Ferber. 2020. Video-based breathing rate monitoring in sleeping subjects. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2458–2464.

[54] A Sapra, A Malik, and P Bhandari. 2023. Vital sign assessment. In: StatPearls. *Treasure Island (FL)* (2023).

[55] Karl Ernst Schaefer. 1958. Respiratory pattern and respiratory response to CO2. *Journal of Applied Physiology* 13, 1 (1958), 1–14.

[56] Ali I Siam, Nirmeen A El-Bahnasawy, Ghada M El Banby, Atef Abou Elazm, and Fathi E Abd El-Samie. 2020. Efficient video-based breathing pattern and respiration rate monitoring for remote health monitoring. *JOSA A* 37, 11 (2020), C118–C124.

[57] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing* 3, 1 (2012), 42–55. https://doi.org/10.1109/T-AFFC.2011.25

[58] Lionel Tarassenko, Mauricio Villarroel, Alessandro Guazzi, Joao Jorge, DA Clifton, and Chris Pugh. 2014. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological measurement* 35, 5 (2014), 807.

[59] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th EC, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 402–419.

[60] Michael J Tipton, Abbi Harper, Julian FR Paton, and Joseph T Costello. 2017. The human ventilatory response to stress: rate or depth? *The Journal of physiology* 595, 17 (2017), 5729–5752.

[61] Andrea Valenzuela, Nicolás Sibuet, Gemma Hornero, and Oscar Casas. 2021. Non-contact video-based assessment of the respiratory function using a rgb-d camera. *Sensors* 21, 16 (2021), 5605.

[62] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. 2013. Phase-based video motion processing. *ACM Transactions on Graphics (ToG)* 32, 4 (2013), 1–10.

[63] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. 2023. Tracking everything everywhere all at once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19795–19806.

[64] Peter Welch. 1967. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics* 15, 2 (1967), 70–73.

[65] Lacey Whited, Muhammad F Hashmi, and Derrel D Graham. 2017. Abnormal respirations. (2017).

[66] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. 2023. PointOdyssey: A Large-Scale Synthetic Dataset for Long-Term Point Tracking. In *ICCV*.