Event-based Visual Vibrometry

Xinyu Zhou[1] Peiqi Duan[2,3] Yeliduosi Xiaokaiti[2,3] Chao Xu[1] Boxin Shi[2,3*]

[1]State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University

[2] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

[3]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

{zhouxiny, duanqi0001, shiboxin}@pku.edu.cn, yongqiye@stu.pku.edu.cn, xuchao@cis.pku.edu.cn

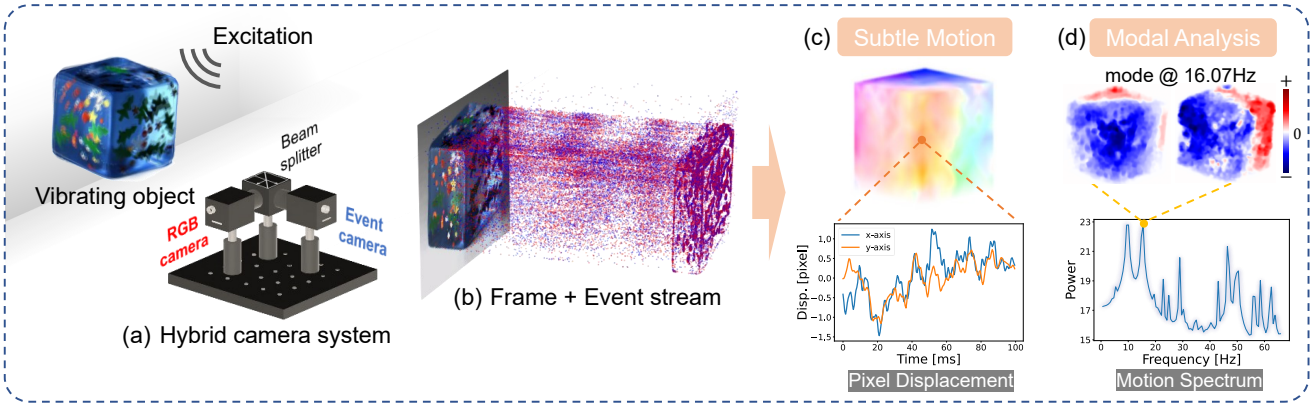


Figure 1. Overview of event-based visual vibrometry. A hybrid camera system, comprising a frame-based camera and an event camera, is utilized to capture subtle object vibrations (a). We propose an event-based subtle motion estimation method that leverages the event stream and a reference frame as input (b). The subtle motion field is extracted at each pixel location across time (c). This motion field can be further analyzed to extract the global motion spectrum and independent vibration modes (d). Each mode is characterized by a resonant frequency and a corresponding mode shape, which manifest as distinct peaks in the motion spectrum.

Abstract

Visual vibrometry has emerged as a powerful technique for remote acquisition of audio and the physical properties of materials. To capture high-frequency vibrations, frame-based approaches often require a high-speed video camera and bright lighting to compensate for the short exposure time. In this paper, we introduce event-based visual vibrometry, a new high-speed visual vibration sensing method using an event camera. By leveraging the high temporal resolution and low bandwidth characteristics of event cameras, event-based visual vibrometry enables high-speed vibration sensing under ambient lighting conditions with improved data efficiency. Specifically, we leverage a hybrid camera system and propose an event-based subtle motion estimation framework that integrates an optimization-based approach based on the event generation model and a motion refinement network. We demonstrate our method by capturing vibration caused by audio sources and estimating material properties for various objects.

* Corresponding author

1. Introduction

Vibrations are pervasive phenomena that arise from a wide range of sources, including engines, percussive impacts, and musical instruments. These vibrations typically exhibit small amplitudes and a broad spectrum of frequencies, spanning from a few Hz to MHz, and carry an immense amount of information. Vibration analysis is an established tool in various engineering and scientific fields, including the remote acquisition of audio signals [8], the monitoring of human heart rate [38, 56], and the estimation of material properties [2, 9, 13].

Various devices have been used to measure and analyze vibrations. Among these, contact sensors [48] and laser vibrometers [11] are two widely used traditional vibration sensors, both of which are specialized instruments and only measure the vibration of a single surface point. Recently, much progress has been made on *visual vibrometry* [2, 4, 9, 13], an emerging technique that utilizes general-purpose video cameras for vibration measurement. In contrast to traditional vibration sensors, video-based approaches enable spatially dense measurements of surface motion. However, the acquisition of high-frequency vi-

brations requires high-speed cameras coupled with bright lighting to compensate for the short exposure time, which is impractical in many real-world applications. Additionally, high-speed cameras demand substantial bandwidth requirements. To eliminate the need for high-speed cameras, Sheinin *et al*. [43] propose a dual-shutter system combined with active speckle-based techniques. Nevertheless, the nature of speckle-based vibration imaging restricts its applicability and confines vibration sampling to only several surface points. In short, frame-based visual vibrometry encounters challenges in efficiently capturing high-frequency vibrations while maintaining bandwidth efficiency and measurement convenience.

Event cameras [14, 25], characterized by their high temporal resolution, high dynamic range, and low bandwidth requirements, have emerged as a promising solution for high-speed sensing [16]. Unlike conventional cameras, event cameras asynchronously record per-pixel logarithmic scene radiance changes. This signal-triggering mechanism is particularly advantageous for vibration sensing, where radiance changes occur only in a small fraction of pixels, leading to a significantly reduced data rate. The high temporal resolution and dynamic range inherent to event cameras enable the measurement of high-frequency vibrations under ambient lighting. These characteristics position event cameras as potential alternatives to address the constraints of frame-based visual vibrometry.

Despite the potential of event cameras, estimating subtle motion from event data remains an open problem. Firstly, although event cameras have been applied to small motion scenarios for motion magnification [7], this approach focuses on visualizing small motions and adopts an implicit motion representation. However, the explicit estimation of subtle motion is essential in visual vibrometry for vibration analysis. Secondly, adapting existing event-based motion estimation techniques to vibration scenarios presents significant difficulties. Existing event-based optical flow estimation methods [19, 37, 47, 58] generally assume continuous apparent motion of visual patterns, which is violated by the oscillatory nature of vibrations. Events caused by vibrations are primarily localized along object edges, and this spatial sparsity hinders the effectiveness of event alignment techniques like Contrast Maximization [37, 47]. Moreover, in comparison to general scenarios, vibrations tend to produce significantly fewer events. Estimating subtle motion from such sparse event data is highly vulnerable to noise inherent in event cameras [23].

In this paper, we propose event-based visual vibrometry as an efficient and practical solution for vibration measurement and analysis. To address the challenges mentioned above, we employ an Event-RGB hybrid camera system [7, 20, 55] (Fig. 1(a)), which combines event streams with spatial texture information from conventional frames. We propose an event-based subtle motion estimation framework that takes event streams and a single reference frame as inputs (Fig. 1(b)). Initially, we employ an optimization-based approach grounded in the event generation model [21, 36] to derive coarse motion estimates within short temporal windows. The sparsity of event data restricts the validity of these coarse estimates to only event-triggered regions, while simultaneously introducing undesirable smoothing effects. Additionally, quantization errors in event cameras compromise the accuracy of motion amplitude estimation. To address these issues, we adopt a motion refinement network that aggregates historical motion information to enhance the coarse motion estimation and obtain long-range vibration displacements (Fig. 1(c)). The refined motion fields across time can subsequently be analyzed (Fig. 1(d)) for various visual vibrometry applications. The key contributions of our work are summarized as follows:

- a bandwidth-efficient visual vibrometry approach that leverages a hybrid camera system to measure high-frequency vibration under ambient lighting conditions;
- an event-based subtle motion estimation framework that integrates a model-based optimization approach and a motion refinement network;
- a thorough evaluation on both synthetic data and real-world vibrations caused by various sources, demonstrating the effectiveness of event-based visual vibrometry.

2. Related Works

Traditional vibration sensors. This section reviews two prevalent traditional vibration measurement devices: contact sensors and active sensors. Contact sensors operate through direct physical contact with the target object [48]. Accelerometers and piezoelectric pickups are representative examples of contact sensors that respond to acceleration. Among active sensors, laser Doppler vibrometers (LDVs) are the most common type used for vibration measurement [11, 41]. LDVs operate by projecting a laser beam onto the target object and calculating the surface velocity from the phase shift in the reflected light. LDVs have been used to find defects in composite materials [3, 5], examine the integrity of building structures [40], and modal analysis [30, 31]. Both contact sensors and laser vibrometers are typically limited to measuring vibrations at a single point on the surface.

Frame-based visual vibrometry. Recent studies on video motion magnification [53, 54, 56] have demonstrated the capability to detect subtle motions captured in videos. These methods typically exploit spatial phase variations of the complex steerable pyramid [39, 49, 50] to estimate small local motions. Chen *et al*. [6] applied this phase-based subtle motion estimation method to quantify the vibration modes of pipes and cantilever beams. Davis

et al. [8] analyzed object vibrations induced by acoustic excitation to reconstruct the original sound. Davis *et al*. [9] employed the motion spectra extracted from video to estimate the physical properties of objects. Feng *et al*. [13] extended this methodology to heterogeneous properties and proposed a physics-constrained optimization approach. Capturing high-frequency vibrations requires expensive 2D high-speed cameras. To address this cost, some works [1, 57] utilize fast 1D sensors for high-frequency acquisition, but they can only measure vibrations along one dimension. Sheinin *et al*. [43] proposed a dual-shutter system with high-frequency sensing capabilities, integrated with active speckle-based techniques that optically magnify small-amplitude vibrations through laser illumination.

Event-based vibration measurement. Recent studies have investigated the application of event cameras for vibration measurement [10, 29, 33]. Dorn *et al*. [10] adapted the phase-based subtle motion extraction method to event data, assuming that each event contributes an equal phase change. Subsequently, they developed an event-based structural monitoring method. Shi *et al*. [44] incorporated laser active illumination to enhance event-based vibration measurements and introduced a frequency estimation technique based on Gaussian mixture distributions. Niwa *et al*. [34] presented a non-contact audio recovery setup using an event camera, employing the phase-based approach proposed in [10]. Howard *et al*. [22] implemented a similar setup, but recovered audio signals by analyzing the zero-crossings of pixels. Chen *et al*. [7] developed an event-based motion magnification method for amplifying and visualizing subtle motions. However, this approach does not directly estimate subtle motions; instead, motion information is implicitly derived from event signals via neural network processing. In this work, we focus on passive capture and estimation of subtle vibrations through the synergistic use of event signals and conventional frames.

3. Method

3.1. Problem Formulation

Vibration modal analysis. The vibrations of objects are often well-approximated by linear systems. In modal analysis, a solid object is modeled as a system of point masses interconnected by springs [42]. In this discretized model, the mass matrix $\mathbf{M}$ and the stiffness matrix $\mathbf{K}$ describe mass concentrated and stiffness between each pair of degrees of freedom (DOF), respectively. The equation of motion for this system is given by:

$$\mathbf{M}\ddot{x} + \mathbf{K}x = 0, \tag{1}$$

where x and $\ddot{x}$ denote the displacement and acceleration of each DOF. The object's vibrations exhibit resonant frequencies, which correspond to the eigenvalues of the system:

$$\mathbf{K}\mathbf{u}_i = \omega_i^2 \mathbf{M}\mathbf{u}_i, \tag{2}$$

where ω_i represents a resonant frequency and the corresponding eigenvector $\mathbf{u}_i$ describes the mode shape. To make the theory of vibration accessible to the computer vision audience, we highlight the following key concepts:

- **Mode Shapes** ($\mathbf{u}_1 ... \mathbf{u}_n$): Each mode shape represents a distinct pattern of vibration for the object. The set of mode shapes forms an orthogonal basis for the vibration.
- **Resonant Frequencies** ($\omega_1 ... \omega_n$): Each mode is associated with a resonant frequency. Resonant frequencies are spatially invariant across the object's surface.

The global motion power spectrum for an object can be computed by averaging the power spectra of local motions extracted at each pixel location. The modes of vibrations manifest as peaks in the motion power spectrum (Fig. 1(d)). Estimating mechanical properties from vibrations is based on the insight that variations in mechanical properties induce changes in the resonant frequencies and mode shapes for a fixed geometry.

Event formation model. When the logarithmic change of brightness at pixel $\mathbf{x}_k$ and time t exceeds a specific threshold C, an event signal $e_k = (\mathbf{x}_k, t_k, p_k)$ will be triggered:

$$|\mathbf{L}(\mathbf{x}_k, t_k) - \mathbf{L}(\mathbf{x}_k, t_k - \Delta t_k)| \geq C, \tag{3}$$

where $p_k \in \{1, -1\}$ is the polarity of the brightness change, $\mathbf{L}(\mathbf{x}_k, t_k)$ denotes the logarithmic brightness at pixel $\mathbf{x}_k$ and time t_k, and $t_k - \Delta t_k$ is the time at which the preceding event was triggered at pixel $\mathbf{x}_k$. Let $\mathcal{E} = \{e_k\}_{k=1}^{N_e}$ denote a set of N_e events triggered within a short time interval, the brightness change can be computed by summing their polarities:

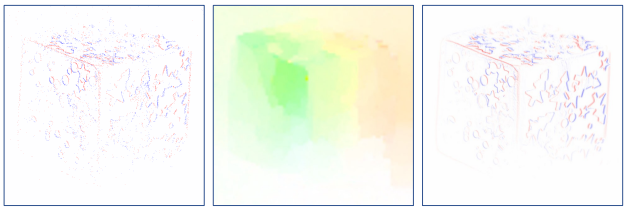
$$\Delta\mathbf{L}(\mathbf{x}) = \sum_{k=1}^{N_e} p_k C \delta(\mathbf{x} - \mathbf{x}_k), \tag{4}$$

where the Dirac delta function δ selects the pixel $\mathbf{x}_k$. Assuming brightness constancy over the interval, the brightness change can be approximated as the product of the brightness gradient $\nabla\mathbf{L}$ and the displacement $\Delta\mathbf{x}$, where the displacement is a result of motion with velocity $\mathbf{v}$ on the image plane [18]:

$$\Delta\widehat{\mathbf{L}}(\mathbf{x}) = -\nabla\mathbf{L}(\mathbf{x}) \cdot \Delta\mathbf{x} = -\nabla\mathbf{L}(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x})\Delta t. \tag{5}$$

3.2. Coarse Motion by Optimization

Given the sparsity of event signals in vibration scenarios, event alignment and correlation techniques [19, 37, 47], commonly used in previous methods, are not applicable.



(a) Events ($\Delta\mathbf{L}$) (b) Coarse Motion (c) Predicted ($\Delta\hat{\mathbf{L}}$)

Figure 2. An example (corresponding to the jelly cube in Fig. 1(b)) of coarse motion through optimization. The predicted brightness change (c), denoted as $\Delta\widehat{\mathbf{L}}$ in Eq. (5) aligns well with the event frame (a), represented by $\Delta\mathbf{L}$ in Eq. (4).

Therefore, we employ a hybrid Event-RGB camera system, integrating high-speed event signals with spatially dense texture information from images, to facilitate subtle motion estimation. Our method takes event streams and a single reference frame as inputs.

Leveraging the event formation model, we formulate the problem of event-based subtle motion estimation as an optimization task, where we minimize the mismatch between the brightness change $\Delta\mathbf{L}$, derived from event accumulation (as shown in Eq. (4)), and its prediction $\Delta\widehat{\mathbf{L}}$, obtained from moving brightness edges (as detailed in Eq. (5)).

Optimization objective. The composite loss function is a combination of data-fidelity term E_{data} and regularization term E_{reg}:

$$E(\mathbf{v}) = E_{\text{data}}(\mathbf{v};\mathcal{E}) + \lambda E_{\text{reg}}(\nabla\mathbf{v}), \tag{6}$$

where λ serves as a balancing parameter, empirically set to 0.5 in our experiments.

The data-fidelity term E_{data} measures the difference between the observed brightness change $\Delta\mathbf{L}$ derived from event accumulation and its prediction $\Delta\widehat{\mathbf{L}}$ based on our coarse motion estimation $\mathbf{v}$:

$$E_{\text{data}}(\mathbf{v};\mathcal{E}) = \left\| \frac{\Delta\widehat{\mathbf{L}}}{\|\Delta\widehat{\mathbf{L}}\|_2} \cdot \mathbf{M} - \frac{\Delta\mathbf{L}}{\|\Delta\mathbf{L}\|_2} \right\|_1, \tag{7}$$

where $\mathbf{M}$ is a mask selecting pixels with events triggered. Due to spatial-temporal variations in the contrast threshold C [23], we compute the difference between normalized brightness changes across the image plane to enhance the robustness of the optimization, following established methodologies [21, 46]. To analyze high-frequency vibrations, the subtle motion is estimated within a very short time interval (typically less than 0.5 ms), during which very few event signals are triggered. It is challenging to accurately estimate flow in regions devoid of events, which often correspond to areas with low contrast or zero flow. To mitigate these inaccuracies, we only calculate the data-fidelity loss term at pixel locations with events triggered.

We assume that flow vectors vary smoothly with sparse discontinuities occurring at object boundaries. The regularization term E_{reg} encourages smoothness of the estimated motion $\mathbf{v}$ by minimizing the differences between neighboring pixels. This term is defined as:

$$E_{\text{reg}}(\nabla\mathbf{v}) = \|\omega(\mathbf{x})\nabla\mathbf{v}(\mathbf{x})\|_1, \tag{8}$$

where ω represents the smoothness weight for each pixel. Following prior research [36], the weight ω is computed based on the gradients of the reference image:

$$\omega(\mathbf{x}) = \exp(-\mu\,|\nabla\mathbf{L}(\mathbf{x})|), \tag{9}$$

where $\nabla\mathbf{L}(\mathbf{x})$ denotes the spatial gradient of the logarithmic brightness at pixel $\mathbf{x}$.

Optimization details. To enhance the convergence speed and robustness of the optimization process, we employ a patch-based optimization strategy implemented in a coarse-to-fine manner [47]. This strategy involves three resolution scales in a pyramid structure, with the coarsest patch size at 32×32 pixels and the finest at 8×8 pixels. Bilinear interpolation is utilized for upsampling between successive scales. $\mathbf{v}$ is initialized uniformly within the range $[-0.1, 0.1]$ at the coarsest scale. For the subsequent scale, the initialization is derived from the optimization results of the preceding scale. We utilize the Adam optimizer [24] with a learning rate of 0.05, and the overall number of iterations is set to 40.

Analysis. An example of coarse motion estimation is presented in Fig. 2. From this example, it is observed that the predicted brightness change $\Delta\widehat{\mathbf{L}}$, derived from the coarse motion (Fig. 2(b)), aligns well with the event frame (Fig. 2(a)), which corresponds to $\Delta\mathbf{L}$ in Eq. (4). This demonstrates the efficacy of the coarse motion estimation. However, the coarse motion results manifest several limitations. First, due to the sparsity of events, the estimated coarse motion yields valid results only in regions where events are triggered, and it is susceptible to noise. Second, quantization errors in event cameras hinder the accurate recovery of motion amplitude in the coarse motion results.

3.3. Motion Refinement by Learning

With the optimization-based approach described in Sec. 3.2, we obtain the coarse motion results, denoted as f^c for ease of presentation. While f^c enables the brightness change prediction $\Delta\widehat{\mathbf{L}}$ to align with the event signals, it exhibits issues including excessive smoothness, susceptibility to noise, and inaccurate motion amplitude estimation. To address these problems, we propose **EVibNet**, a motion refinement network specifically designed to enhance the coarse motion results.

The pipeline of EVibNet is shown in Fig. 3. The network takes event signals $\mathcal{E}$, a single reference frame $\mathbf{I}_{\text{ref}}$,

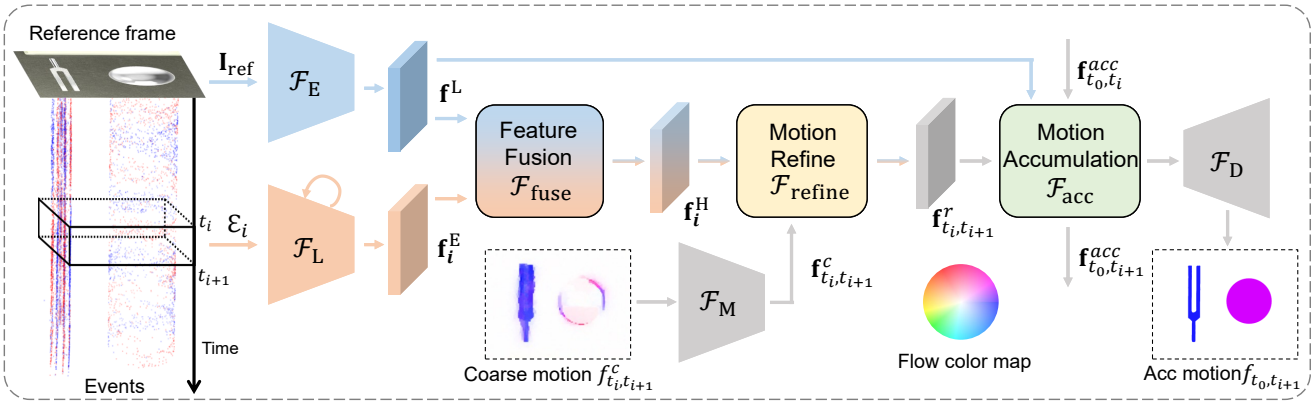


Figure 3. Pipeline of our EVibNet. (**Input:** $\mathbf{I}_{\text{ref}}$ and $\mathcal{E}_i$) Initially, features are extracted from the reference frame $\mathbf{I}_{\text{ref}}$ and events triggered within the time interval $[t_i, t_{i+1}]$ using $\mathcal{F}_{\text{L}}$ and $\mathcal{F}_{\text{E}}$, respectively. Then the extracted features are fused with a fusion module $\mathcal{F}_{\text{fuse}}$. The coarse motion obtained in Sec. 3.2 is fed into a motion encoder $\mathcal{F}_{\text{M}}$, generating the motion feature $\mathbf{f}^{c}_{t_i,t_{i+1}}$, which is further refined by the motion refine module $\mathcal{F}_{\text{refine}}$. Finally, the refined motion feature is accumulated to obtain the vibration displacement (**Outout:** $\mathbf{f}_{t_0,t_{i+1}}$).

and coarse motion results f^c as inputs. Specifically, EVibNet comprises two principal components: 1) two modality-specific encoders $\mathcal{F}_{\text{L}}$ and $\mathcal{F}_{\text{E}}$ and a fusion module $\mathcal{F}_{\text{fuse}}$; 2) a coarse motion refinement module $\mathcal{F}_{\text{refine}}$ and a motion accumulation module $\mathcal{F}_{\text{acc}}$.

Feature extraction. For events $\mathcal{E}_i$ triggered in the time interval $[t_i, t_{i+1}]$, we transform $\mathcal{E}_i$ into two distinct voxel grids $\mathbf{E}^+_i$ and $\mathbf{E}^-_i$ for positive and negative events respectively, following the procedure in [17]. Due to the modality gap between frames and events, we use two modality-specific encoders $\mathcal{F}_{\text{L}}$ and $\mathcal{F}_{\text{E}}$ to extract features from the reference frame $\mathbf{I}_{\text{ref}}$ and events $\mathcal{E}_i$, respectively. To solve the issue of inaccurate motion amplitude estimation caused by quantization error, we leverage the temporal structure inherent in the event data using a recurrent convolutional encoder. Specifically, a ConvLSTM [45] layer is integrated into each encoder stage of $\mathcal{F}_{\text{E}}$. The feature extraction process for the reference frame and event stack is performed using the feature encoders as follows:

$$\mathbf{f}^{\text{L}} = \mathcal{F}_{\text{L}}(\mathbf{I}_{\text{ref}}), \quad \mathbf{f}^{\text{E}}_i = \mathcal{F}_{\text{E}}(\mathbf{E}^+_i, \mathbf{E}^-_i, h_{i-1}), \tag{10}$$

where $\mathbf{f}^{\text{L}}, \mathbf{f}^{\text{E}}_i \in \mathbb{R}^{h \times w \times D}$, the feature resolution (h, w) is 1/4 of the original image resolution (H, W), and h_{i-1} refers to the hidden states in the recurrent event encoder.

Given that event signals are triggered in only a few regions when capturing vibrating objects, it is crucial to leverage texture features extracted from the reference frame to regulate the motion features extracted from the events. Following [51], we employ a cross-modal attention fusion block, denoted as $\mathcal{F}_{\text{fuse}}$, to adaptively fuse the information from these two modalities, yielding the fused feature $\mathbf{f}^{\text{H}}_i$. The cross-modal attention in $\mathcal{F}_{\text{fuse}}$ takes as input the queries $\mathbf{Q}^{\text{E}}$ derived from the event features $\mathbf{f}^{\text{E}}_i$, and the keys $\mathbf{K}^{\text{L}}$ and values $\mathbf{V}^{\text{L}}$ derived from the image features $\mathbf{f}^{\text{L}}$.

Motion refinement and accumulation. With the frame-event fused feature $\mathbf{f}^{\text{H}}_i$, we refine the coarse motion $f^c_{t_i,t_{i+1}}$ through a motion refinement module, denoted as $\mathcal{F}_{\text{refine}}$. The coarse motion $f^c_{t_i,t_{i+1}}$ is initially encoded into coarse motion features $\mathbf{f}^c_{t_i,t_{i+1}}$ using a motion encoder $\mathcal{F}_{\text{M}}$. Subsequently, the motion information embedded in $\mathbf{f}^{\text{H}}_i$ is leveraged to refine $\mathbf{f}^c_{t_i,t_{i+1}}$, producing the refined motion feature $\mathbf{f}^r_{t_i,t_{i+1}}$:

$$\mathbf{f}^r_{t_i,t_{i+1}} = \mathcal{F}_{\text{refine}}(\mathbf{f}^{\text{H}}_i, \mathbf{f}^c_{t_i,t_{i+1}}). \tag{11}$$

Following the refinement of coarse motion, we further accumulate short-term motion into long-range displacement. This approach is motivated by two factors: First, motion within a very short time interval is typically minimal and thus more susceptible to noise. Second, estimating the displacement of points on vibrating objects relative to a reference state facilitates subsequent analysis. However, direct accumulation of motion at each time step results in significant accumulation errors. Therefore, we integrate short-term motion using a motion accumulation module $\mathcal{F}_{\text{acc}}$. The accumulated motion feature $\mathbf{f}_{t_0,t_i}$ is computed as follows:

$$\mathbf{f}^{acc}_{t_0,t_{i+1}} = \mathcal{F}_{\text{acc}}(\mathbf{f}^{acc}_{t_0,t_i}, \mathbf{f}^r_{t_i,t_{i+1}}, \mathbf{f}^{\text{L}}), \tag{12}$$

where $\mathbf{f}^{acc}_{t_0,t_i}$ is derived from the preceding time interval, and $\mathbf{f}^{\text{L}}$ serves as the context feature to guide the motion accumulation. For simplicity, $\mathcal{F}_{\text{acc}}$ is implemented as a simple RNN comprising two ConvNext blocks [27]. Finally, the accumulated motion feature is decoded into the target displacement $f_{t_0,t_{i+1}}$ using a motion decoder $\mathcal{F}_{\text{D}}$.

3.4. Implementation Details

Loss functions. Our loss functions are derived from prior work on optical flow estimation [52]. The network is supervised using the l_1 distance between the predicted and

ground truth flow across the entire sequence of predictions, $\{f_{t_0,t_1}, ..., f_{t_0,t_N}\}$. Given the ground truth flow sequence $\{f^{gt}_{t_0,t_1}, ..., f^{gt}_{t_0,t_N}\}$, the loss is defined as:

$$\mathcal{L} = \sum_{i=1}^{N} \| f^{gt}_{t_0,t_i} - f_{t_0,t_i} \|_1. \tag{13}$$

Synthetic dataset. Obtaining ground truth data for real-world subtle motion is challenging. Therefore, following previous studies [7, 35], we simulate subtle motion by compositing foreground elements onto a background. We sample images from the MS COCO dataset [26] for the background and use segmented objects from the PASCAL VOC dataset [12] for the foreground. The magnitude and direction of motion for both the background and each foreground object are generated randomly. To simulate sub-pixel motions, we initially generate high-resolution videos, where the motion appears larger, and subsequently downsample each frame to the desired resolution. After obtaining the videos, we use the event simulator V2E [23] to simulate event data from downsampled videos.

Training details. We implement our approach using the PyTorch framework and conduct all experiments on a single NVIDIA GeForce RTX 3090 GPU. During training, we use the AdamW optimizer [28] and configure the batch size to 4. The EVibNet is trained together for 20 epochs. The initial learning rate is set to 2×10^{-4}, and a cosine annealing learning rate scheduling strategy is employed.

Hybrid camera system. Our hybrid camera system consists of a machine vision camera (HIKVISION MV-CA050-12UC) and an event camera (PROPHESEE GEN4.1), which are co-aligned using a beam splitter (Thorlabs CCM1-BS013). We utilize the calibration method proposed in [32] to achieve pixel alignment.

4. Experiments

In this section, we first evaluate the sub-pixel motion estimation accuracy of the proposed method on the synthetic dataset in Sec. 4.1. Then we validate our method's ability to detect high-frequency vibrations by capturing and replaying the vibration caused by audio sources in Sec. 4.2. Finally, we demonstrate our method by estimating both homogeneous and heterogeneous material properties of objects with known geometry in Sec. 4.3.

4.1. Evaluation on Synthetic Dataset

We compare our method to the phase-based subtle motion estimation approach (denoted as RGBPhase) that is widely used in frame-based visual vibrometry. This approach computes local motion signals from phase shifts within a complex steerable pyramid (CSP) [39, 49, 50]. For the comparison with high-speed videos, we apply RGBPhase to the original video sequences used to simulate event streams in our synthetic dataset. Furthermore, we compare our method with two representative event-based optical flow estimation methods: Multi-reference Contrast Maximization (MCM) [47] and the recurrent version of EV-FlowNet [37, 58]. To validate the superior suitability of our method for subtle motion estimation compared to the event-based motion magnification method (EMM) [7], we retrain EMM using our synthetic dataset and leverage an off-the-shelf optical flow estimation method RAFT [52] to estimate optical flow from the motion-magnified videos generated by EMM.

Table 1. Motion estimation comparison on the synthetic dataset. **E** and **F** denote events and frames, respectively. ↓ indicates the lower, the better throughout this paper. The best performances are highlighted in **bold**.

Methods	Inputs	EPE ↓	% Out ↓
EV-FlowNet [37]	**E**	0.1805	47.95
MCM [47]	**E**	0.4511	66.87
RGBPhase [15]	**F** (960Hz)	0.2833	54.78
EMM [7]+RAFT [52]	**E** + **F** (30Hz)	0.2190	47.63
Ours	**E** + **F** (single)	**0.0834**	**21.29**

We adopt the average endpoint error (EPE) and the percentage of pixels with EPE $>$ 0.1 px (denoted as "% Out") as evaluation metrics. The quantitative comparison results are presented in Tab. 1. The results demonstrate that our method substantially outperforms previous event-based optical flow estimation techniques, primarily due to its specific design for subtle motion estimation. Notably, RGBPhase exhibits unsatisfactory performance on our synthetic dataset, likely attributable to the noise in local phase information encoded in the CSP. The inferior performance of EMM + RAFT compared to our method suggests that subtle motion estimation cannot be effectively achieved by simply cascading an optical flow method with EMM. This performance gap may be attributed to the blurring artifacts often introduced by motion magnification techniques.

4.2. Capturing audio signals

To validate the efficacy of our method in real-world high-frequency vibration sensing, we apply it to remote acquisition of audio signals. In the experimental setup depicted in Fig. 4, we capture the membrane vibrations of a speaker playing a linear up-chirp signal ranging from 0 to 1000 Hz. Concurrently, we obtain reference audio recordings using a microphone. With the subtle motion extracted by our method, the 1D audio signal is reconstructed by averaging the local motions across all pixel locations and orientations [8]. We compare our method with an event-based sound recovery method (EBVM) proposed by Niwa *et al.* [34]. By integrating events and frames, our method estimates subtle motion more accurately than the phase-based ap-

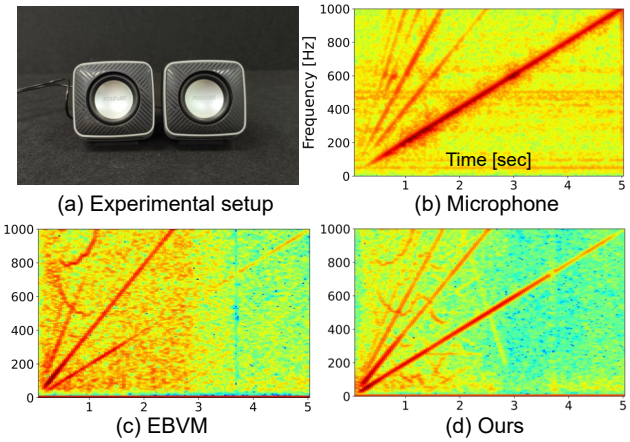


Figure 4. Speaker membrane experiments. (a) A speaker is playing an up-chirp signal ranging from 0 to 1000 Hz. (b) The spectrogram of the source sound recorded by a microphone. (c) The spectrogram of the recovered audio from EBVM [34]. (d) The spectrogram of our recovered sound.

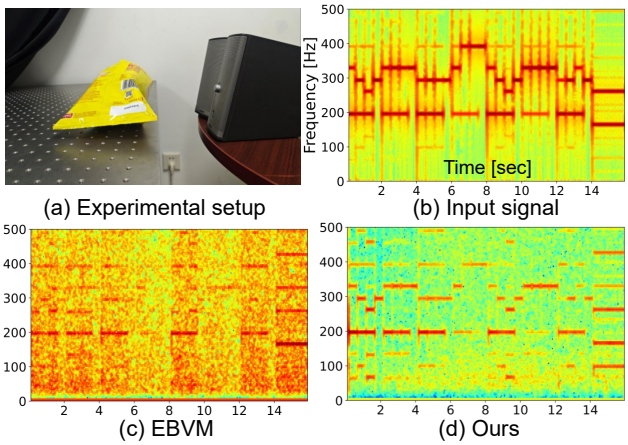


Figure 5. Audio recovery from a bag of chips. (a) Sound from an audio source (speaker) excites the chip bag. (b) The spectrogram of the input signal sent to the speaker. (c) The spectrogram of the recovered audio from EBVM [34]. (d) The spectrogram of our recovered sound.

Table 2. Percent error in estimating the Young's Modulus for each rod. The best performances are highlighted in **bold**. The content in each cell refers to the results for rods clamped to a length of 35 cm and 48 cm respectively.

% Error ↓	Lighting	Aluminum	Steel	Copper
Video [9]	low	10.01/10.42	-11.64/-9.28	-9.40/4.64
	medium	4.13/-3.70	-10.72/-7.62	-4.41/-2.70
	high	-1.45/-1.99	-7.94/-5.96	-3.14/-2.10
Ours	low	2.90/3.21	-7.61/-6.58	-5.65/2.28
	medium	-1.26/1.64	-6.77/-5.08	-3.39/-2.07
	high	**0.39/-1.47**	**-5.93/-5.08**	**-2.25/-2.07**

proach adopted in EBVM. The experimental results demonstrate significantly reduced noise levels and enhanced high-frequency reconstruction fidelity compared to the baseline method.

In Fig. 5, we conduct an experiment similar to Davis *et al*. [8]. A speaker playing the "MaryMIDI" audio excites the chip bag. We measure the vibration of a chip bag and recover the audio. The spectrograms of recovered audio reveal superior waveform fidelity achieved by our method. These results demonstrate our method's robustness and practical applicability in complex real-world vibration sensing scenarios.

4.3. Estimating material properties

We apply our method to estimate homogeneous and heterogeneous material properties from vibrations.

Homogeneous material properties. Considering objects with homogeneous material properties, it can be obtained from Eq. (2) that different objects with the same geometry have an identical set of mode shapes, but their resonant frequencies scale proportionally to material properties [9]. Following Davis *et al*. [9], we apply our method to estimate the homogeneous material properties of various metal rods. The simple geometry of clamped rods makes their vibration well-studied [42]. The fundamental frequency ω_1 (first resonant frequency) of a rod is given by:

$$\omega_1 = 0.1399 \frac{d}{L^2} \sqrt{\frac{E}{\rho}}, \tag{14}$$

where d is the diameter of the rod, L is the length, E is the Young's modulus, and ρ is the density. While length, diameter, and density can be easily measured, Young's modulus is often measured with a tensile test, which will damage the object. In this experiment, we try to estimate the Young's modulus of the rods by finding their fundamental frequency, which is the highest peak in the motion spectrum.

We test rods with three different metals: aluminum, steel, and copper. Each rod is tested at two lengths (35 cm and 48 cm) and under three brightness conditions (100, 500, and 1500 lux). For comparison with frame-based visual vibrometry, we capture 1000 fps videos under identical settings. Resonant frequencies are global properties and invariant to the viewpoint, ensuring that estimation accuracy is not affected by the viewpoint. In Tab. 2, we compare the estimated Young's modulus values with those provided by the manufacturer. The results demonstrate that our method achieves more robust performance across varying lighting conditions. We further compare the data sizes of 1000 fps videos with signals captured by our imaging system in the rod experiments. Our method requires only a single image; consequently, the data size of the reference frame is negli-

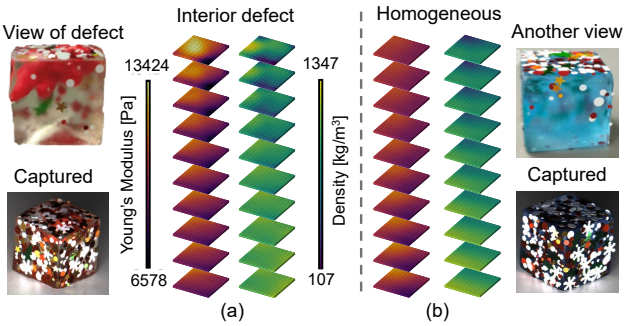


Figure 6. Heterogenous material properties estimation result of the defect cube (a) and a homogeneous cube (b), the results are plotted with the same colormaps. Based on the results, we can easily differentiate between a cube with a defect and a homogeneous one.

gible. At the same 720p resolution, event cameras produce event streams at 1-5 MB/s, about 20% of the compressed 1000 fps videos (H. 264 encoded). The discrepancy is more pronounced when considering the actual transmission bandwidth required for high-speed cameras. We also test existing event-based optical flow estimation methods. However, both EV-FlowNet [37] and MCM [47] struggle to accurately estimate the vibrations of the rods, and we can not identify reasonable resonant frequencies in the motion spectra.

Heterogeneous material properties. When considering objects with heterogeneous material properties, spatial inhomogeneities in these properties influence both the resonant frequencies and mode shapes. Based on the mathematical relationship between modes and material properties, Feng *et al*. [13] propose a physical-constrained optimization approach to estimate heterogeneous material properties from several image-space vibration modes.

To demonstrate the applicability of our method to visual vibration tomography, we experiment on two jelly cubes: one homogeneous and the other with an interior defect (a rubber doll embedded within the jelly cube). We capture data under an initial deformation condition and identify unique modes in the estimated motion fields, as shown in Fig. 1. For each cube, we infer spatially-varying Young's modulus and density values on a $10 \times 10 \times 10$ hexahedral mesh with the optimization algorithm proposed by Feng *et al*. [13]. The reconstruction results are shown in Fig. 6. The reconstruction is obtained using only five unique image-space modes. The defective cube has a rubber doll on top, which is stiffer than the jelly cube. The results Fig. 6(a) show the existence of the defect, and clearly differ from that of the homogeneous cube. Due to the high damping of the jelly cubes, the reconstruction results are not perfectly aligned with the data, but we can still easily differentiate between the two cubes.

Table 3. Ablation study of the method design on the synthetic data.

	Coarse	W/o cm	W/o acc	W/o frame	Ours
EPE ↓	0.3632	0.0969	0.1299	0.1734	**0.0834**
% Out ↓	51.45	25.91	36.99	42.25	**21.29**

4.4. Ablation studies

To validate the effectiveness and necessity of each part of our method, we conduct several ablation studies and show the results in Tab. 3. To assess the necessity of motion refinement using EVibNet, we test the performance of coarse motion (denoted as "coarse"). To verify the contribution of the coarse motion obtained from optimization, we remove it from the inputs of EVibNet (denoted as "w/o cm"). To demonstrate the effectiveness of the motion accumulation module, we directly accumulate motions at each time step (denoted as "w/o acc"). Finally, we show the necessity of the reference frame by removing it from inputs (denoted as "w/o frame"). As indicated in Tab. 3, our complete method achieves the lowest EPE and % Out, which demonstrates the contribution of each component of our method.

5. Conclusion

We present a bandwidth-efficient approach for high-speed vibration sensing with an Event-RGB hybrid camera system. To extract subtle motion frame event data, we integrate an optimization-based approach with a learning-based motion refinement network. We demonstrate our method through the capture of vibrations induced by audio sources and the estimation of material properties.

Limitations. In the synthetic dataset, subtle motion is simulated by compositing foreground elements onto a background, which exhibits a gap from real-world vibration scenarios. In the future, simulating more realistic synthetic datasets using the finite element method (FEM) will further improve the performance of event-based visual vibrometry.

Acknowledgement

This work was supported by National Natural Science Foundation of China (Grant No. 62088102, 62136001, 62402014, 62276007), Beijing Natural Science Foundation (Grant No. L233024), and Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012). Peiqi Duan was also supported by China National Postdoctoral Program for Innovative Talents (Grant No. BX20230010) and China Postdoctoral Science Foundation (Grant No. 2023M740076). The authors thank openbayes.com for providing computing resource.

References

[1] Silvio Bianchi and Emanuele Giacomozzi. Long-range detection of acoustic vibrations by speckle tracking. *Applied optics*, 58(28):7805–7809, 2019. 3

[2] Katherine L Bouman, Bei Xiao, Peter Battaglia, and William T Freeman. Estimating the material properties of fabric from video. In *Proc. of IEEE International Conference on Computer Vision*, 2013. 1

[3] Oral Büyüköztürk, Mehmet Ali Taşdemir, O Büyüköztürk, R Haupt, C Tuakta, and J Chen. Remote detection of debonding in frp-strengthened concrete structures using acoustic-laser technique. In *Proc. of Nondestructive Testing of Materials and Structures*, 2013. 2

[4] Oral Buyukozturk, Justin G Chen, Neal Wadhwa, Abe Davis, Frédo Durand, and William T Freeman. Smaller than the eye can see: Vibration analysis with video cameras. In *Proc. of World Conference on Non-Destructive Testing*, 2016. 1

[5] Justin G Chen, Robert W Haupt, and Oral Buyukozturk. The acoustic-laser vibrometry technique for the noncontact detection of discontinuities in fiber reinforced polymer-retrofitted concrete. *Materials evaluation*, 72(10), 2014. 2

[6] Justin G Chen, Neal Wadhwa, Young-Jin Cha, Frédo Durand, William T Freeman, and Oral Buyukozturk. Modal identification of simple structures with high-speed video using motion magnification. *Journal of Sound and Vibration*, 345:58–71, 2015. 2

[7] Yutian Chen, Shi Guo, Fangzheng Yu, Feng Zhang, Jinwei Gu, and Tianfan Xue. Event-based motion magnification. In *Proc. of European Conference on Computer Vision*, 2024. 2, 3, 6

[8] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J. Mysore, Frédo Durand, and William T. Freeman. The visual microphone: passive recovery of sound from video. *ACM Transactions on Graphics*, 33(4):79:1–79:10, 2014. 1, 3, 6, 7

[9] Abe Davis, Katherine L Bouman, Justin G Chen, Michael Rubinstein, Oral Büyüköztürk, Frédo Durand, and William T Freeman. Visual Vibrometry: Estimating material properties from small motions in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):732–745, 2017. 1, 3, 7

[10] Charles Dorn, Sudeep Dasari, Yongchao Yang, Garrett Kenyon, Paul Welch, and David Mascareñas. Efficient full-field operational modal analysis using neuromorphic event-based imaging. In *Proc. of Conference and Exposition on Structural Dynamics*, 2017. 3

[11] Franz Durst, Adrian Melling, and James H Whitelaw. Principles and practice of laser-doppler anemometry. *NASA STI/Recon Technical Report A*, 76:47019, 1976. 1, 2

[12] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 6

[13] Berthy T Feng, Alexander C Ogren, Chiara Daraio, and Katherine L Bouman. Visual vibration tomography: Estimating interior material properties from monocular video. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3, 8

[14] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Pooria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, et al. A 1280× 720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86 μm pixels, 1.066 geps readout, programmable event-rate controller and compressive data-formatting pipeline. In *Proc. of IEEE International Solid-State Circuits Conference*, 2020. 2

[15] David J Fleet and Allan D Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5:77–104, 1990. 6

[16] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 2

[17] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-End learning of representations for asynchronous event-based data. In *Proc. of IEEE International Conference on Computer Vision*, 2019. 5

[18] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. EKLT: asynchronous photometric feature tracking using events and frames. *International Journal of Computer Vision*, 128(3):601–618, 2020. 3

[19] Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense continuous-time optical flow from event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4736–4746, 2024. 2, 3

[20] Jin Han, Yixin Yang, Peiqi Duan, Chu Zhou, Lei Ma, Chao Xu, Tiejun Huang, Imari Sato, and Boxin Shi. Hybrid high dynamic range imaging fusing neuromorphic and conventional images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8553–8565, 2023. 2

[21] Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 4

[22] Matthew Howard and Keigo Hirakawa. Event-based visual microphone. In *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2023. 3

[23] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbrück. V2E: From video frames to realistic DVS events. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2, 4, 6

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of International Conference on Learning Representations*, 2015. 4

[25] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43 (2):566–576, 2008. 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. of European Conference on Computer Vision*, 2014. 6

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 5

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. of International Conference on Learning Representations*, 2017. 6

[29] Yuanyuan Lv, Liang Zhou, Zhaohui Liu, and Haiyang Zhang. Structural vibration frequency monitoring based on event camera. *Measurement Science and Technology*, 35(8): 085007, 2024. 3

[30] William N MacPherson, Mark Reeves, David P Towers, Andrew J Moore, Julian DC Jones, Martin Dale, and Craig Edwards. Multipoint laser vibrometer for modal analysis. *Applied optics*, 46(16):3126–3132, 2007. 2

[31] M Martarelli, Gian Marco Revel, and C Santolini. Automated modal analysis by scanning laser vibrometry: problems and uncertainties associated with the scanning system calibration. *Mechanical systems and signal processing*, 15 (3):581–601, 2001. 2

[32] Manasi Muglikar, Mathias Gehrig, Daniel Gehrig, and Davide Scaramuzza. How to calibrate your event camera. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 6

[33] Woong-jae Na, Kyung Ho Sun, Byeong Chan Jeon, Jaeyun Lee, and Yun-ho Shin. Event-based micro vibration measurement using phase correlation template matching with event filter optimization. *Measurement*, 215:112867, 2023. 3

[34] Ryogo Niwa, Tatsuki Fushimi, Kenta Yamamoto, and Yoichi Ochiai. Live demonstration: Event-based visual microphone. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 3, 6, 7

[35] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed A. Elgharib, Frédo Durand, William T. Freeman, and Wojciech Matusik. Learning-based video motion magnification. In *Proc. of European Conference on Computer Vision*, 2018. 6

[36] Liyuan Pan, Miaomiao Liu, and Richard Hartley. Single image optical flow estimation with an event camera. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2, 4

[37] Federico Paredes-Vallés, Kirk YW Scheper, Christophe De Wagter, and Guido CHE De Croon. Taming contrast maximization for learning sequential, low-latency, event-based optical flow. In *Proc. of IEEE International Conference on Computer Vision*, 2023. 2, 3, 6, 8

[38] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762–10774, 2010. 1

[39] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40:49–70, 2000. 2, 6

[40] Nicolaas Bernardus Roozen, Ludovic Labelle, Monika Rychtáriková, and Christ Glorieux. Determining radiated sound power of building structures by means of laser doppler vibrometry. *Journal of Sound and Vibration*, 346:81–99, 2015. 2

[41] Steve Rothberg, JR Baker, and Neil A Halliwell. Laser vibrometry: pseudo-vibrations. *Journal of Sound and Vibration*, 135(3):516–522, 1989. 2

[42] Ahmed A Shabana. *Theory of vibration*. Springer, 1991. 3, 7

[43] Mark Sheinin, Dorian Chan, Matthew O'Toole, and Srinivasa G Narasimhan. Dual-shutter optical vibration sensing. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3

[44] Chenyang Shi, Ningfang Song, Boyi Wei, Yuzhen Li, Yibo Zhang, Wenzhuo Li, and Jing Jin. Event-based vibration frequency measurement with laser-assisted illumination based on mixture gaussian distribution. *IEEE Transactions on Instrumentation and Measurement*, 72:1–13, 2023. 3

[45] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, 2015. 5

[46] Shintaro Shiba, Friedhelm Hamann, Yoshimitsu Aoki, and Guillermo Gallego. Event-based background-oriented schlieren. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2011–2026, 2024. 4

[47] Shintaro Shiba, Yannick Klose, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow, depth and ego-motion estimation by contrast maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7742–7759, 2024. 2, 3, 4, 6, 8

[48] Peter J Shull. *Nondestructive evaluation: theory, techniques, and applications*. CRC press, 2002. 1, 2

[49] Eero P Simoncelli and William T Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proc. of International Conference on Image Processing*, 1995. 2, 6

[50] Eero P Simoncelli, William T Freeman, Edward H Adelson, and David J Heeger. Shiftable multiscale transforms. *IEEE transactions on Information Theory*, 38(2):587–607, 1992. 2, 6

[51] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *Proc. of European Conference on Computer Vision*, 2022. 5

[52] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *Proc. of European Conference on Computer Vision*, 2020. 5, 6

[53] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM Transactions on Graphics*, 32(4):1–10, 2013. 2

[54] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Riesz pyramids for fast phase-based video magnification. In *Proc. of International Conference on Computational Photography*, 2014. 2

[55] Zihao W Wang, Peiqi Duan, Oliver Cossairt, Aggelos Katsaggelos, Tiejun Huang, and Boxin Shi. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[56] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics*, 31(4):1–8, 2012. 1, 2

[57] Nan Wu and Shinichiro Haruyama. The 20k samples-per-second real time detection of acoustic vibration based on displacement estimation of one-dimensional laser speckle images. *Sensors*, 21(9):2938, 2021. 3

[58] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. In *Proc. of Robotics: Science and Systems*, 2018. 2, 6

Event-based Visual Vibrometry

Supplementary Material

Xinyu Zhou[1] Peiqi Duan[2,3] Yeliduosi Xiaokaiti[2,3] Chao Xu[1] Boxin Shi[2,3*]

[1]State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University

[2] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

[3]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

{zhouxiny, duanqi0001, shiboxin}@pku.edu.cn, yongqiye@stu.pku.edu.cn, xuchao@cis.pku.edu.cn

6. Additional experimental results

This section presents additional experimental results demonstrating the efficacy of our method in visual vibrometry applications.

6.1. Audio recovery under different lighting

The high dynamic range (HDR) characteristic of event cameras enables event-based visual vibrometry to operate effectively under ambient illumination conditions. To analyze the performance of our method under varying illumination levels, we record one speaker playing a chirp signal at four distinct brightness levels, ranging from 400 lux to 3200 lux. For comparison of the frame-based method, we simultaneously capture the speaker with a high-speed video camera at 1000 fps and recover audio signals using the visual microphone (VM) technique [2]. The signal reconstruction quality is evaluated using the segmental signal-to-noise ratio (SSNR). As shown in Tab. 4, our method achieves more robust performance across different lighting conditions. Notably, the performance of our method under low illumination (400 lux) is comparable to that of the frame-based method under high illumination (3200 lux).

Table 4. Signal reconstruction SSNR (the higher the better) comparison under different lighting conditions.

Method	400 lux	800 lux	1600 lux	3200 lux
VM [2]	0.31	0.38	0.89	2.25
Ours	3.75	4.48	4.51	4.65

6.2. Analyzing vibration of tuning forks

We analyze the vibrations of two tuning forks with fundamental frequencies of 128 Hz and 256 Hz, respectively. As shown in Fig. 7, we strike the forks with a rubber-tipped mallet and measure their vibrations. The spectrograms obtained by our method accurately reflect the fundamental frequencies of the tuning forks. In contrast, EBVM fails to recover these fundamental frequencies, potentially due to noise in the event signal under ambient lighting conditions.

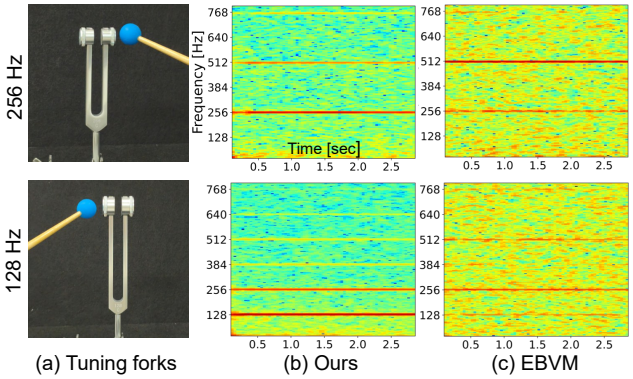


Figure 7. Vibration analysis of two tuning forks with fundamental frequencies of 128Hz and 256Hz (a). We compare recovered spectrograms between our method (b) and EBVM [5] (c). The results obtained using our method align well with the fundamental frequencies of the tuning forks.

6.3. Material properties with unknown geometry

We demonstrate the applicability of our method in learning the material properties of objects with unknown geometry. The experiments on material property estimation, as detailed in the main manuscript, rely on precise knowledge of the object's geometry. As a result, their potential application is limited to objects with simple geometries that can be precisely measured, or to man-made structures with detailed CAD models, for which resonant frequencies can be obtained through the finite element method (FEM).

Given a set of objects with similar but not precisely modeled geometries, the differences in their material properties will be revealed in their resonant frequencies and mode shapes. Based on this intuition, Davis *et al*. propose to learn relationships between motion spectra and the material properties of objects with similar but unknown geometry. They conducted experiments on a dataset comprising 30 hanging fabrics [1], along with corresponding ground truth measurements of area weight. Following their work, we simu-

* Corresponding author

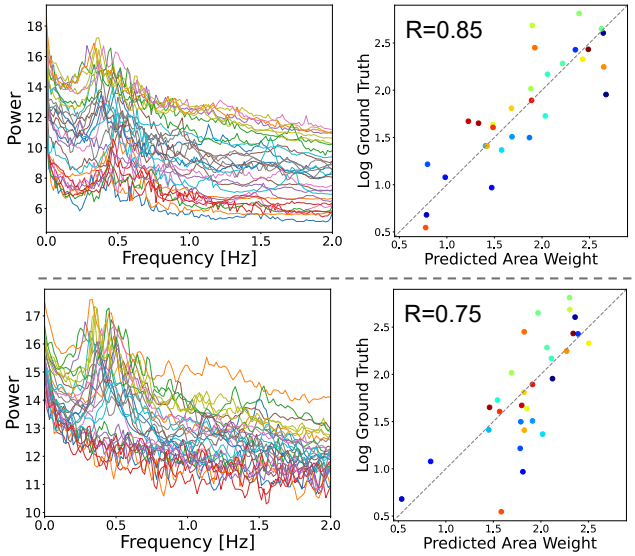


Figure 8. Comparison of extracted motion spectra and predictions on material properties estimated from videos [3] (upper) and events (below) on the fabric dataset [1]. The Pearson correlation values (R) are shown in the figure.

late corresponding events with the V2E simulator [4] from videos captured by a grayscale Point Grey camera at 60 fps. From the simulated events, we extract motion spectra using our proposed method. Consistent with the methodology in [3], we employ the motion spectra directly as features and train a Partial Least Squares Regression model to map the motion spectra to the logarithm of the ground truth area weight. Due to the small size of the dataset, we employ the leave-one-out cross-validation strategy. The extracted motion spectra and area weight prediction results obtained from our method and from videos are presented in Fig. 8. The Pearson correlation values (R) of our predictions are slightly lower than those of the frame-based method. Note that the data size of our simulated events (16-bit Prophesee EVT 3.0 format) is less than 10% that of the original videos, indicating that the vibrations are efficiently encoded in the simulated events. Considering the reduced data size, we believe the minor performance drop is acceptable.

7. Discussion

7.1. Thresholds' impact on vibration sensing

For frame-based cameras, the accuracy of vibration estimation depends on bit depth and quantum efficiency in noise-free conditions. Correspondingly, the precision of subtle motion estimation from event data is affected by the contrast threshold. The threshold of event cameras generally exceeds one gray level in images, resulting in a lower theoretical precision for event-based visual vibrometry, especially in scenarios involving extremely low-amplitude vibration measurements. Intuitively, lowering the threshold would trigger more event signals, thereby improving motion extraction accuracy. Nevertheless, due to current hardware constraints, event cameras are more susceptible to noise at lower thresholds. Despite this hardware limitation, experimental results demonstrate that our method achieves satisfactory performance across many applications. Future hardware advancements will further improve the precision of event-based visual vibrometry.

7.2. Inference frequency

The time step for voxel partitioning is primarily determined by the vibration frequency of the observed object. Specifically, the Nyquist frequency of the motion estimation results should exceed the target frequency range. The computational overhead of our method increases linearly with the number of voxels. At a resolution of 256×256, the overall running time for each step of the coarse motion optimization and subsequent network refinement on our test system is approximately 0.04s.

We evaluate the quality of audio recovered from the sequence capturing a speaker playing the "MarySpeech" audio file used by Davis *et al*. [2] under 4 distinct inference frequencies: 1000, 2000, 4000, 6000Hz. The intelligibility (STOI) scores are [0.481, 0.513, 0.524, 0.523] (higher is better). Increasing the inference frequency expands the detectable vibration spectrum, thereby enhancing signal reconstruction fidelity. However, this concurrently reduces the event signal density per voxel, which may affect the precision of micro-vibration estimation. Correspondingly, the STOI scores exhibit an initial ascent followed by a plateau. In our experiment, we set the inference frequency slightly above the target frequency range.

7.3. Temporal filtering

Previous motion magnification studies typically employ temporal filtering to select motion within specific frequency bands of interest using a band-pass filter. Temporal filtering helps to prevent noise from being magnified, but it requires prior knowledge of the observed vibration. In contrast, our method usually aims to analyze vibrations across a broad frequency spectrum. To mitigate noise in subtle motion estimation, our approach employs a reference image and exploits the temporal structure inherent in event data, which is extracted through a recurrent event encoder, thereby effectively suppressing isolated noise events.

7.4. Dynamic scenes

When the observed object undergoes global motion, it is more challenging to detect subtle vibrations. Following previous studies [2, 3], our method assumes that the observed object remains static, exhibiting only tiny vibrations. Under this assumption, our approach utilizes only one im-

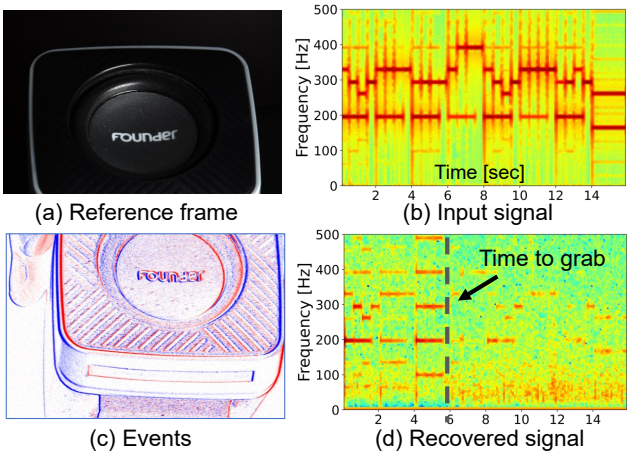


Figure 9. We capture a speaker that is manually grasped and shaken. (a) The reference frame. (b) The spectrogram of the input signal sent to the speaker. (c) Event signals at another timestamp. (d) The spectrogram of our recovered sound.

age to provide scene texture information and is inapplicable to dynamic scenes. To evaluate our method on non-static scenes, we conduct an experiment where a speaker is manually grasped and shaken. As shown in Fig. 9, the spectrogram of the recovered signal reveals a performance drop when the speaker is shaken. Future works could track the object's macro-motion using both frames and events, and dynamically update the reference frame.

References

[1] Katherine L Bouman, Bei Xiao, Peter Battaglia, and William T Freeman. Estimating the material properties of fabric from video. In *Proc. of IEEE International Conference on Computer Vision*, 2013. 1, 2

[2] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J. Mysore, Frédo Durand, and William T. Freeman. The visual microphone: passive recovery of sound from video. *ACM Transactions on Graphics*, 33(4):79:1–79:10, 2014. 1, 2

[3] Abe Davis, Katherine L Bouman, Justin G Chen, Michael Rubinstein, Oral Büyüköztürk, Frédo Durand, and William T Freeman. Visual Vibrometry: Estimating material properties from small motions in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):732–745, 2017. 2

[4] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbrück. V2E: From video frames to realistic DVS events. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[5] Ryogo Niwa, Tatsuki Fushimi, Kenta Yamamoto, and Yoichi Ochiai. Live demonstration: Event-based visual microphone. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1