Learning to See Inside Opaque Liquid Containers using Speckle Vibrometry

Matan Kichler Shai Bagon Mark Sheinin
Weizmann Institute of Science, Israel
matankic@gmail.com, shai.bagon@weizmann.ac.il, mark.sheinin@weizmann.ac.il
https://matankic.github.io/see_inside_containers

Abstract

*Computer vision seeks to infer a wide range of information about objects and events. However, vision systems based on conventional imaging are limited to extracting information only from the visible surfaces of scene objects. For instance, a vision system can detect and identify a Coke can in the scene, but it cannot determine whether the can is full or empty. In this paper, we aim to expand the scope of computer vision to include the novel task of inferring the hidden liquid levels of opaque containers by sensing the tiny vibrations on their surfaces. Our method provides a first-of-a-kind way to inspect the fill level of multiple sealed containers remotely, at once, without needing physical manipulation and manual weighing. First, we propose a novel speckle-based vibration sensing system for simultaneously capturing scene vibrations on a 2D grid of points. We use our system to efficiently and remotely capture a dataset of vibration responses for a variety of everyday liquid containers. Then, we develop a transformer-based approach for analyzing the captured vibrations and classifying the container type and its hidden liquid level at the time of measurement. Our architecture is invariant to the vibration source, yielding correct liquid level estimates for controlled and ambient scene sound sources. Moreover, our model generalizes to unseen container instances within known classes (*e.g.*, training on five Coke cans of a six-pack, testing on a sixth) and fluid levels. We demonstrate our method by recovering liquid levels from various everyday containers.*

1. Introduction

Since its inception in the 1960s, computer vision has sought to enable machines to infer a wide range of information about scene objects and events. In industrial settings, this capability enables the automation of various inspection and monitoring tasks involving manufactured goods, mechanical parts, or stored items across a range of environments. Today, much like in its early days, most computer vision systems rely on sensing using conventional cameras designed to replicate the human eye. However, using conventional cameras limits the scope of retrievable scene information. For instance, an image or video can be used to detect and identify a scene object, but the object's internal properties, like its content or material composition, are mostly indeterminable from its captured surface appearance. In this paper, we focus on revealing one commonplace hidden object property: the fill level of opaque liquid containers.

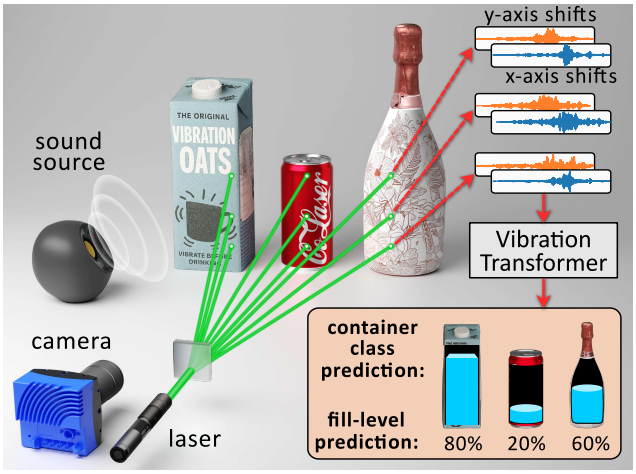


Figure 1. Learning to 'see' the fill level of opaque containers. We develop a novel system to capture the surface vibrations of multiple liquid containers at multiple surface points. We vibrate the containers using a nearby speaker and train a novel Vibration Transformer to infer the container class and the hidden liquid level, expressed as a percentage of total capacity, from the measured multi-point surface vibrations.

Our work joins a rich body of previous vision research focused on retrieving hidden object properties by leveraging *unconventional* imaging. Clever hyperspectral scene probing and polarization cues were used to classify [27, 42–44] and segment [28, 50] scene object materials. Material properties were also probed by thermal imaging [10, 12, 33, 37, 41] and structured light methods [30]. While effective, these methods can only probe object properties that are optically accessible at the object's surface.

To probe deeper, one must use a signal that permeates the object's interior but can be optically sensed at the object's surface. One such signal is object vibrations. In a series of seminal works, Bounman, Davis *et al.*, captured the minute vibrations of simple objects like rods and fabrics using high-speed cameras to extract properties like object density and Young's modulus [6, 13]. Later, Feng *et al.*, extended the vibrometry-based method to infer volumetric stiffness and density [19, 35]. However, capturing high-frequency, low-amplitude vibrations on general object surfaces using high-speed cameras is challenging due to bandwidth and optical magnification constraints. To address these challenges, Sheinin *et al.*, recently demonstrated modal analysis using a dual-shutter camera prototype that leverages laser speckle to capture vibrations at high speeds for colinear scene points [46]. Later, Zhang *et al.*, used the same system to recover the anisotropy of different-material planar objects [58].

The visual vibration works described above focused on recovering low-level object properties like motion spectra, material stiffness, or density. In this paper, we seek to infer a higher-level object semantic property: the amount of liquid it presently holds. Specifically, similar to Davis *et al.* [13], we excite the scene using a nearby speaker and measure the resulting object vibrations using a novel speckle-based imaging system. Our system is inspired by prior speckle-based vibration works [5, 7, 46, 55–57]. However, unlike previous works, it can capture a 2D grid of scene points simultaneously, enabling vibration measurement of multiple objects at once, with each object probed at multiple points on its surface (see Figs. 1 and 3). Notably, several prior works tackled the liquid level recovery task by either recording the sound of liquid pouring [4, 54], or recording the sound resulting from a physical knock on the container [21]. Conversely, our approach eliminates the need for any physical interaction with the container or reliance on nearby microphones, enabling passive, remote inference of liquid content using only visual measurements.

Inferring the hidden liquid level from a container's vibrations is a challenging task. The vibrational response of an object having a simple shape and material composition can be modeled using a small set of material parameters (*e.g.*, mass and stiffness). As shown in prior works, these parameters can be recovered by observing the object's resonant frequencies and mode shapes [13, 19]. While some liquid containers, such as certain wine glasses, exhibit a simple relationship between the fluid level and the resulting resonant frequencies, most everyday containers feature complex geometries and heterogeneous materials, resulting in a non-trivial relationship between the liquid level and the resulting vibrations. The challenge becomes substantial when inferring fluid levels of unseen containers from the same class, as slight manufacturing variations can shift resonant frequencies even among identical-looking items (*e.g.*, cans in a six-pack). To address this challenge, we develop a learning-based approach introducing a novel physics-inspired 'Vibration Transformer' to classify the fluid level of various everyday liquid containers. The Vibration Transformer receives a spectral decomposition of the recorded two-axis, multi-point surface vibrations. It is, thus, invariant to the speaker excitation (content and duration).[1]

Our approach introduces a novel way to remotely assess the fill levels of multiple sealed containers at once without requiring any physical handling or manual weighing. We believe our method could facilitate the inspection of large warehouses storing consumer liquid products (*e.g.*, soda cans, shampoos), heavy containers impractical to weigh, and containers holding hazardous, toxic, flammable, or radioactive liquids (*e.g.*, heavy or tritiated water). The latter, long-storage containers, are especially susceptible to spillage and evaporation over time—issues that are further exacerbated by repeated physical handling and inspection.

To evaluate our approach, we gather a dataset of various everyday liquid containers and measure their vibration response at three surface points per container for multiple speaker positions and short (two-second) excitations (*e.g.*, a logarithmic chirp, a segment of a popular song, and ambient noises). We show that our approach can accurately classify the hidden liquid levels of the dataset containers for novel speaker positions and fluid levels not seen during training. Moreover, we show that if trained on multiple same-class containers (*e.g.*, five Coke cans), the model can infer the hidden liquid levels of containers outside the training set (*i.e.*, a sixth Coke can).

2. Background

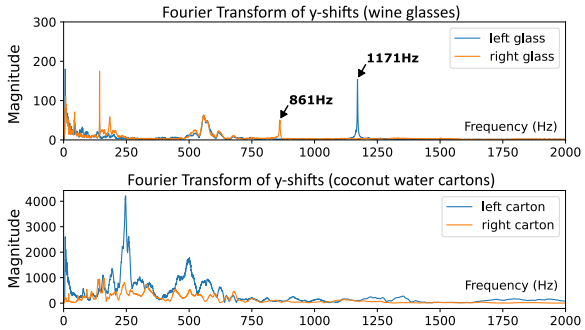
2.1. Modeling object vibrations

Under small deformations, most objects can be approximated as linear elastic, meaning they exhibit a proportional relationship between stress and strain and return to their original shape once the disturbance is removed. The vibrations of objects in this regime can be modeled using a second-order linear differential equation called the *equation of motion*, which depends on the object's material properties (*i.e.*, mass, stiffness, and damping), boundary conditions, and external forces (*i.e.*, the vibration excitation) [9, 14].

In certain conditions, the vibrations of linear elastic objects can be expressed as a linear combination of their *resonant frequencies*, or vibration modes [14]. Measuring these modes—defined by a set of independent natural frequencies and mode shapes—enables inference of the equation of motion parameters mentioned above [13, 19]. For everyday liquid containers, the equation of motion must account for the coupled interaction between the container and the liquid inside. The presence of liquid introduces added

[1] As long as the excitation signal is sufficiently broadband.

(a) single point experiment



(b) Fourier transforms of y-axis shifts

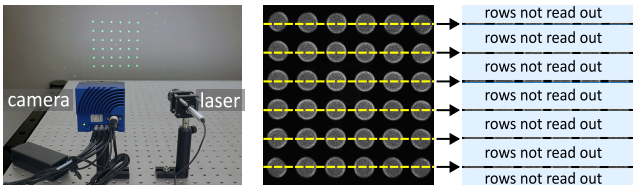
Figure 2. Variations in container responses. **(a)** Setup showing vibration measurements from a resonant wine glass (clear for illustration) and a standard beverage container. We record responses from two identical items at different fill levels using a single surface point per object. **(b)** The wine glass exhibits a distinct resonance that shifts with fill level. In contrast, the coconut water container exhibits a complex frequency response that is difficult to relate to its fill level. We train a transformer to infer fill levels from such signals by analyzing vibrations at *multiple* surface points.

mass, increasing inertia and lowering resonant frequencies (see Fig. 2). Additionally, fluid-structure interactions alter stiffness and damping, resulting in complex vibrations with a non-trivial relationship to the container's liquid level. Furthermore, slight manufacturing differences in the same container type may yield different resonance frequencies even for the same fluid level (Fig. 7(a)). Therefore, in this work, we use a learning-based approach to infer this complex relation, for various container types and instances.

2.2. Measuring surface vibrations

Surface vibrations can be measured using various principles, including high-speed imaging [9, 13, 19], interferometry (*e.g.*, Laser Doppler Vibrometers (LDV)) [39], contact-based piezoelectric transducers [29], and more. In this paper, we rely on speckle-based vibrometry, which consists of illuminating an object's surface points with a laser and capturing a defocused video of the illuminated point [47, 57].

Due to the laser light coherency, the defocused laser spot will contain a random interference pattern called *speckle* [1, 57]. This speckle pattern is highly sensitive to small surface tilts, causing it to shift within the point's defocused spot [46, 57]. Thus, unlike LDVs, speckle-based vibrometry obviates the need for expensive specialized hardware and simplifies vibration sensing to computing image-domain shifts. In Sec. 3, we describe a novel vibration-sensing method allowing vibration capture for a 2D grid of surface points.



(a) simplified prototype **(b)** full resolution **(c)** defined ROIs

Figure 3. Capturing vibrations on a 2D grid. **(a)** Our system consists of a laser projecting a 2D grid of points on the scene, and a single defocused camera. The camera image, shown in (b), captures the speckle interference for all grid points. **(b)** We define a set of regions of interest (ROIs) centered on the middle of each defocused row of laser points (ROI centers marked in yellow), with each ROI having a height of a few pixels (*e.g.*, six). **(c)** The camera outputs only the concatenated ROIs, enabling high sampling rates of speckle vibrations (e.g., 57 kHz for six ROIs of 6 pixels each).

3. Sensing rapid vibrations on a 2D grid

Our vibration sensing system measures vibrations for a 2D grid of scene surface points. As such, it enables simultaneous multi-point vibration sensing for multiple scene containers. For simplicity, Sec. 3 and 4 focus on a single container, though the same method is independently applied to each scene container. Our system operates on a simple yet effective principle, capturing speckle vibrations robustly at high speeds in previously undemonstrated configurations.

As shown in Fig. 3(a), our prototype consists of a single laser and a camera. The laser illuminates the scene through a custom diffractive beam splitter that splits the beam into a 2D grid (*e.g.*, 6×6). The split beam then passes through an anamorphic prism pair, configured to widen the beam angles along the horizontal axis, thereby aligning the laser points with the scene's container arrangement (see Fig. 5).

The camera is defocused to yield a grid of speckle patches (Fig. 3(b)). Recovering the two-axis vibrations at each sensed laser grid point involves computing the image-domain shifts of the speckle within each point's patch [57]. However, at full resolution, the camera's bandwidth severely limits the maximum frame rate, resulting in insufficiently fast vibration sampling rates. Therefore, to achieve high sampling rates, we configure the camera's readout to output only M regions of interest (ROIs) of size $W \times P$ pixels, which speeds up the camera's FPS by approximately a factor of $H/(MP)$, where H and W are the camera's full image resolution height and width, respectively. For example, the camera in Fig. 3 can operate at 2247 Hz at full resolution. However, for $M = 6$ ROIs of height $P = 6$ pixels, the camera FPS jumps to 57 699 Hz, a rate sufficient for most mechanical and acoustic vibration applications. See

Sec. 5 for full hardware details and Sec. 7 for a discussion relating our system to prior works.

To robustly compute the two-axis image-domain shifts, $v \in \mathbb{R}^2$, between every two consecutive frames, we develop an ad-hoc method that first uses phase-correlation (PC) [26, 36] to recover integer-pixel shifts, followed by a Lukas-Kanade (LK) [31] estimation of the residual sub-pixel translation (similar to [17]). To handle the high volume of shift computations for recovering multiple laser points at high frame rates (*e.g.*, 1.44 million calls for 2 sec at 20 kHz with a 6×6 grid), we implement a parallelized batched GPU version of PCLK, called PCLK+, to streamline vibration recovery. Our GPU implementation is $\times 20$ faster.

Let $\boldsymbol{v}_i \in \mathbb{R}^{2\times N}$ denote the two-axis image-domain speckle shifts recovered by our camera at surface point i, where $i \in \{0, 1, 2\}$ and N is the number of time samples at camera sampling rate f_{cam} [Hz].[2] The measured signal $\boldsymbol{v}_i$ has units of pixels and can be converted to surface tilts by multiplying with a per-axis scalar [46]. However, since our method does not require such calibration, for simplicity, we will refer to $\boldsymbol{v}_i$ as the object *vibrations* at point i. Next, we describe how to extract the container's fill level given the set of measured container vibrations $\{\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2\}$.

4. Learning to infer a container's liquid level

We aim to recover the container's liquid level from raw vibration signals $\{\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_2\}$ recorded at three surface points. Here, we describe the Vibration Transformer – a transformer-based model that takes the multi-point vibration signals $\boldsymbol{v}_i$ and outputs the liquid level as a percentage of the container's total capacity (*e.g.*, 20 % full).

While prevailing signal processing frameworks operate in the temporal (*e.g.* [38, 51]) or the short time Fourier transform (STFT) domains (*e.g.* [2, 8, 18, 22]), motivated by classic modal analysis, where an object's vibrational modes are recovered to infer low-level object material characteristics, we operate solely in the Fourier domain. That is, the input to our model is

$$V_i\,[f] \equiv |\mathcal{F}\{\boldsymbol{v}_i\}\,| \in \mathbb{R}^2, f \in F^{\text{fixed}} \tag{1}$$

where $\mathcal{F}\{\}$ denotes the Discrete Fourier Transform, f is frequency, and F^{fixed} is a predefined frequency set (*e.g.*, $F^{\text{fixed}} = \{40, 41, 42, .., 2500\}$ [Hz]). Note that the definition of Eq. (1) allows for $\boldsymbol{v}_i$ of an arbitrary duration, content and sampling frequency $f_{\text{samp}} > 2\max(F^{\text{fixed}})$.

The complexity of regressing the multi-point multi-axis Fourier response $V_i\,[f]\,, i \in \{0, 1, 2\}$, to the container's liquid level can vary greatly between different container types. In rare cases, a container's Fourier analysis can easily relate to the liquid level. For example, some wine glasses have a highly resonant response, yielding a distinct high-frequency audible tone that predictably decreases in frequency as the glass is filled (see Fig. 2(a)). However, none of the containers in our dataset have such simple characteristics. Rather, we found that most everyday containers exhibit complex spectral responses that are difficult to interpret and regress using classical methods (*e.g.*, Fig. 2(b)). Therefore, we base our model on the popular Transformer architecture [52], letting the model learn the non-trivial relation between the vibrations at multiple points and the liquid level.

Further inspired by modal analysis, our Vibration Transformer comprises two main components. In the first stage, a shared *PointTransformer* independently processes the signal from each point, analyzing the frequency responses and local resonance characteristics (*i.e.*, mode frequencies). Then, a subsequent *ShapeTransformer* fuses the information from these local features, enabling the model to reason about the mode shapes. Ultimately, the Vibration Transformer's latent representation captures both liquid level and container type from the raw vibration input.

Vibration Transformer architecture. A schematic of our model is shown in Fig. 4. The frequency representation for each point V_i is divided into fixed-sized non-overlapping tokens via a Tokenizer module, where each token encodes information about a specific frequency band Δf. A learnable positional encoding is added to each token, and the sequence of input frequency tokens is supplemented with a trainable [pnt] token [15]. The PointTransformer uses self-attention to extract information from all frequency bands into the [pnt] token per point i, independently.

Then, the ShapeTransformer processes the sequence of three 'transformed' tokens $[\text{pnt}]_i$. Here, additional positional encoding is added to each token $[\text{pnt}]_i$, encoding the grid position of point i on the container. A trainable [cls] token is added to integrate the information from all measured points. We use two MLPs, one to infer the container type, represented as a discrete label c, and another for the discrete liquid level l. In this work, we set $l \in L = \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$.

We train our model using supervised learning. Each collected training sample includes $\{\boldsymbol{v}_0(t)\,, \boldsymbol{v}_1(t)\,, \boldsymbol{v}_2(t)\,; c, l\}$. Our network produces two probability distribution vectors: one for the container class $\hat{c} \in \mathbb{R}^{K_{\text{cont}}}$ and one for the liquid fill level $\hat{p} \in \mathbb{R}^6$. Let $h = 1, 2, .., 6$ denote the index within vector $\hat{p}$ and the corresponding index within the ordered set L. In contrast to the container class prediction, which we optimize using a standard cross-entropy loss, the inherently ordinal nature of the liquid fill level leads us to employ a variation of the Sorted ORDinals (SORD) Loss [16, 20, 40]. For a given true fill level $l \in L$, we define a soft target distribution vector $q_l \in \mathbb{R}^6$, where

$$q_l[h] = \left(e^{-50(l-L[h])^2}\right) / \sum_j e^{-50(l-L[j])^2}, \tag{2}$$

[2] WLOG, we assume each container is illuminated with 3 laser points.

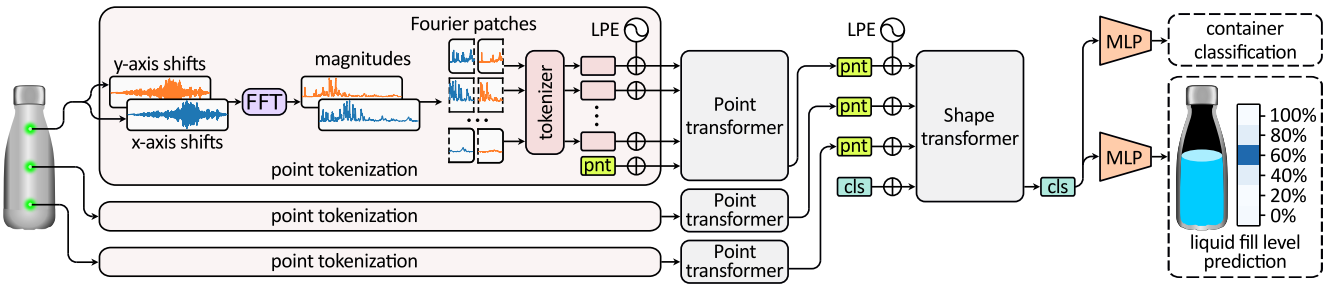


Figure 4. Vibration Transformer network architecture. Our model processes the signals after conversion to the Fourier domain. The signal from each surface point is processed separately by a shared PointTransformer that extracts information from the spectrum of each point. Then, the added [pnt] tokens produced by each PointTransformer are processed by a second, ShapeTransformer which integrates the information between all the surface points. Finally, the ShapeTransformer output token is passed through two MLPs where one classifies the container type and the other the liquid fill level. Here LPE stands for 'learned positional encoding'.

The SORD Loss is then given by:

$$\mathcal{L}_{\text{SORD}} = -q_l^T \cdot \log(\hat{p}) \,. \tag{3}$$

The overall loss is a weighted sum of the cross-entropy loss for c and the SORD Loss for l.

We use the output logits vector $\hat{p}$ to predict the liquid level in two ways. We use a Maximum a posteriori (MAP) estimator to predict the most likely discrete liquid level

$$\hat{l}_{\text{MAP}} \equiv L[h_{\text{MAP}}], \quad h_{\text{MAP}} = \arg\max_h \hat{p}[h]. \tag{4}$$

Additionally, we wish to test the model's behavior for liquid levels in between the levels of L. In such cases, we compute the expectation over the prediction

$$\hat{l}_{\mathbb{E}} = \sum_h L[h] \cdot \hat{p}[h] \tag{5}$$

Note that the model is trained to predict the discrete levels $\hat{l}_{\text{MAP}}$, while $\hat{l}_{\mathbb{E}}$ can be used to output values $\in [0, 1]$.

5. Implementation details

5.1. Hardware details

Our system comprises an EoSens2.0MCX12 camera [32] and a Coherent Sapphire 532 nm 500 mW laser [11]. The laser is passed through a HOLO-OR beamsplitter, yielding a 6×6 point grid with a separation angle of 2.75×2.75 degrees [23]. Thus, at 500 mW, each point has a 13.9 mW power. To illuminate a row of six containers at once, we pass the laser grid through an unpaired Thorlabs anamorphic prism pair [49]. The pair is adjusted to create the desired horizontal spread, turning the square laser grid into a rectangular one.

To probe the row of containers, we create an array of six Creative Pebble V2 speakers [48], which can be individually activated via a ten-channel MOTU UltraLite-mk5 audio interface [34]. The speakers are mounted on a frame detached from the containers, which sit on a vibration-isolated optical breadboard, thus minimizing the transition of mechanical vibrations through the table (see Fig. 5(a)).

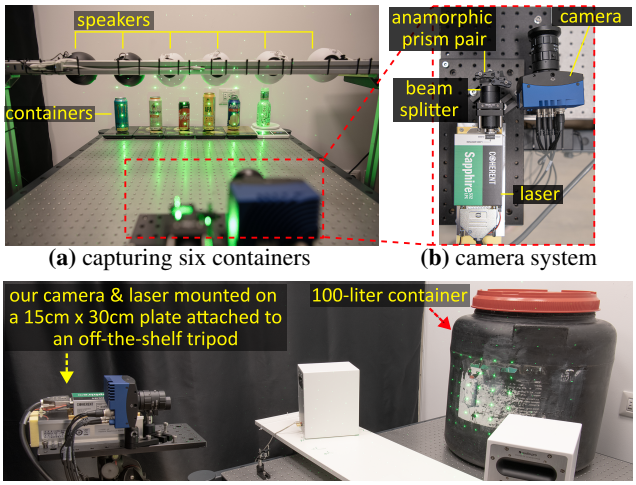
5.2. Training the Vibration Transformer

First, we compute the DFT magnitude, per axis, for each vibration signal $\boldsymbol{v}_i$ on the predefined frequency set $F^{\text{fixed}} = \{100, 100.5, \ldots, 2500\,\text{Hz}\}$. The resulting V_i is a 2×4800 matrix of Fourier magnitude coefficients. We then divide this matrix into non-overlapping patches of size 2×100 (*i.e.*, each patch contains 200 coefficients). For each patch, we apply a linear projection to map the 200-dimensional coefficient vector into a 512-dimensional token. This produces a sequence of 48 tokens, which are subsequently fed into our PointTransformer. We add learnable position encoding to the tokens. Our PointTransformer and ShapeTransformer have eight transformer layers each, with four heads in the self-attention layers. There are about 25 million trainable parameters in each transformer. The prediction MLPs have one hidden layer with 64 dimensions and ReLU activation.

We use the Adam [25] optimizer with a learning rate $\lambda{=}10^{-5}$ to train the model for 7500 epochs. During training, we augment the data by simulating random smooth frequency responses that modulate the magnitudes of the Fourier coefficients (*i.e.*, filters) – mimicking variations in speaker type and acoustic environmental characteristics – and randomly drop 50 % of the PointTransformer input tokens. We give $\mathcal{L}_{\text{SORD}}$ a weight of 0.9 and 0.1 for $\mathcal{L}_{\text{CE}}$.

6. Experimental evaluation

To gather data for training and testing our model, we built an experimental rig comprised of our camera, a set of six speakers positioned on a beam above the test table, and a set of six kitchen scales to measure the amount of liquid we add to each container (see Fig. 5). In each data sampling iteration, we place a set of six containers on the scales filled to one of the liquid levels in $L^{\text{standard}} = \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. Then, we continuously capture their vibrations at three surface points per container while playing multiple short audio sequences using one or more of the six speakers. While the cam-



(a) capturing six containers **(b)** camera system

(c) large container experiment

Figure 5. Experimental setup. **(a)** We acquire vibrations for multiple containers, at multiple points per container at once, for various fill levels. The containers are excited by six different speakers positioned on a beam separated from the container set to avoid transferring mechanical vibrations. Our setup allows simulating sound coming from various directions and amplitudes. We use scales to measure the added liquid level. **(b)** Vibration sensing system. **(c)** We additionally test our system on a large industrial container weighing about 100 kg when full.

era can capture at much higher speeds, we set its frame rate to 5100Hz, having observed that most containers have little energy at frequencies higher than 2000Hz. A subset of containers were captured with extra fill level $L^{\text{interm}} = \{0.25, 0.50, 0.75\}$.

In addition to the small-sized everyday containers described above, our dataset also includes one large 100-liter industrial container (see Fig. 5(c)). This container has a 0.53-meter diameter and weighs about 100 kg when full, and thus cannot be easily placed on any off-the-shelf scale. In total, our dataset contained 5910 individual data samples. Please see the supplementary materials for a full dataset breakdown and visualization.

We partitioned the container dataset into subsets to test the model's prediction for several increasingly complex inference tasks. We train our model on 23 unique container types to predict the levels in L^{standard}, using two types of excitations: a two-second logarithmic chirp with start and end frequencies of 100 and 2500Hz, respectively, and a two-second segment of a popular song. Then, we test our model in six validation categories: (a) *within-distribution*, (b) *unseen instances*, (c) *unseen liquid levels*, (d) *ambient sound*, (e) *unseen levels at ambient sound* and (f) *unseen instances at ambient sound*. [3]

[3] Since the container of Fig. 5(c) is a single unique instance, it was only included in tests *(a)* and *(d)*. See the supplementary for more details.

Test name	Level Pred. Acc. ↑	Level Pred. MAE ↓	Container Acc. ↑
(a) within distribution	0.98	0.01 ±0.03	1.00
CNN baseline	0.17	0.33 ±0.22	0.86
(b) unseen instances	0.79	0.09 ±0.19	0.95
(c) unseen liq. levels	N/A	0.12 ±0.11	0.81
(d) ambient sound	0.92	0.04 ±0.12	0.97
(e) unseen liq. levels + ambient sound	N/A	0.15 ±0.15	0.67
(f) unseen instances + ambient sound	0.59	0.16 ±0.23	0.77

Table 1. Experimental validation results. MAE denotes the mean absolute error. Liquid levels in our settings are bounded between 0 (empty) and 1.0 (full). Thus, an MAE of *e.g.* 0.01 is equivalent to 1% error (chance ≈ 30%). Container classification accuracy is computed over 23 classes, while level accuracy treats level prediction as classification into the predefined set L^{standard} (not applicable to unseen levels). Test scenarios are presented in order of increasing difficulty, where each subsequent test after (a) requires the model to generalize to additional factors not present in the training set. See the supplementary for detailed results.

Test (a): within-distribution. Here we test the model's ability to predict the fill level of containers included in the training set, but for novel speaker positions never seen by the model during training. This corresponds to a task of measuring the liquid level of non-disposable containers that have been previously 'seen' by the system and require repeated testing (*e.g.*, in factories or offices where fill levels must be monitored regularly). To effect this test, we randomly exclude one of the six speakers for each of the training set containers. This designates about 20% of our data for testing. Tab. 1 shows that our model excels at this task, yielding only 1% mean absolute error (MAE) on the test set.

Test (b): unseen instances. Here we test the model on five new containers, each similar in type to those in the training set but not seen during training (*e.g.*, training on five cans of a six-pack, and testing on the sixth). The train set contained three to seven examples for each tested container. Our model achieves good predictions, with 9% error over the test set. Notably, as shown in Fig. 7, we found there could be manufacturing variations between seemingly identical containers, yielding a deviation in the spectral responses. Thus, for category (b), we expect that more training examples should better capture class statistics and improve performance.

Test (c): unseen liquid levels. Here, we test how well our model generalizes to liquid levels outside the training distribution. Despite only being trained to predict levels in L^{standard}, we hypothesized that the ordinal nature of our prediction would lead to reasonable predictions outside L^{standard}. To test this category, we applied the model to con-

Test name	Full model	Single point
(a) within distribution	0.02 ±0.05	0.03 ±0.07
(b) unseen instances	0.09 ±0.15	0.11 ±0.19
(f) unseen instances + ambient sound	0.16 ±0.21	0.18 ±0.21

Table 2. Ablation comparing the full model to a variant that uses only a single surface point measurement.

tainers with L^{interm}, predicting the fill level using the $\hat{l}_{\mathbb{E}}$ estimator (Eq. (5)). Results show 12% test error.

Test (d): ambient sound. Our model was trained on containers *actively* excited by a nearby speaker. However, most environments already contain ambient noises. In this category, we tested the model by playing ambient supermarket background noise. Simulating the ambient sounds using speakers was necessary because our lab is quiet by design. Evaluation on this unseen, structureless excitation yields good predictions (4% error), suggesting our Fourier-based approach is largely invariant to the excitation audio.

Lastly, for completion, tests (e) and (f) contain additional combinations of cases (b),(c), and (d) with increasingly complex inference tasks. Nevertheless, our model performs reasonably, even under these edge scenarios (*e.g.*, unseen liquid levels tested using ambient sound). Overall, our model provides good liquid level predictions despite being trained on a relatively small dataset. Therefore, we believe that increasing the dataset by orders of magnitude would improve all the tests in Tab. 1.

CNN-based baseline. We tested several naive approaches before adopting the transformer architecture in Fig. 4. Notably, we implemented an eight-layer convolutional neural network (CNN) applied independently to each $\boldsymbol{v}_i$, followed by eight more layers on the concatenated hidden representations. Each layer uses a kernel size of 15, BatchNorm [24], and ReLU activation. The first three layers in each component also perform stride-2 average pooling with a 3x-sized kernel. Overall, the CNN had 27M parameters in the first component and 40M in the second, roughly matching the Vibration Transformer's parameter count. Tab. 1 shows this approach classifies container types reasonably well but fails to recover liquid levels, performing no better than chance.

Discrete vs. continuous liquid level prediction. Replacing classification over L levels trained with $\mathcal{L}_{\text{SORD}}$ by a single continuous output trained with ℓ_1 loss raises the MAE from 0.01 to 0.20 on the within-distribution test.

Does phase play a pivotal role? The input to our network is the signal *magnitude*, discarding phase information (Eq. (1)). We explored versions where we retain the phase by either concatenating it to the magnitude or by processing the raw complex $\mathcal{F}\{\boldsymbol{v}_i\}$. Neither performed better than our base model. See the supplementary for detailed results.

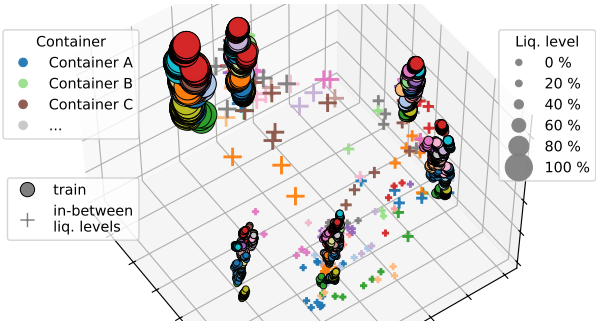


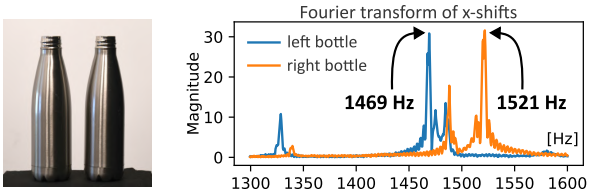
Figure 6. We explore the learned latent space by projecting [cls] token representations using PCA. Marker size reflects liquid level; color indicates container type. Solid markers (training samples) form distinct fill-level clusters with smooth container-type transitions along the z-axis. Faded markers ("unseen liq. levels" test set) appear interpolated between clusters.

How many points? Our prototype collects three vibration measurements per container. To evaluate the impact of multi-point data, we trained models using only one measurement while having the same number of transformer layers. Tab. 2 shows that while a single point suffices in a within-distribution setting, harder cases – like unseen instances of known containers – can benefit from multi-point data which encapsulates the mode shapes.

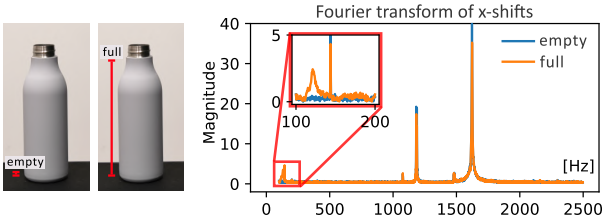
Exploring the learned latent space. Fig. 6 visualizes our model's internal representation of the input data. The plot is generated by extracting the [cls] tokens from all the training samples and applying PCA to project these embeddings onto three principal components. The result reveals six distinct, elongated vertical clusters matching the six discrete liquid fill levels present in our training set. Within each cluster, container types are not randomly scattered; instead, they exhibit an organized shift along the z-axis, with a gradual transition from one container type to another. The plot also contains the samples from the "unseen liquid levels" test (displayed as faded markers). Their position between clusters suggests that, despite training on discrete levels, the model learns a latent representation capturing the continuous spectrum of liquid fill levels. The clear structure shown in Fig. 6 suggests the model learned meaningful features rather than overfitting to individual training examples, despite our dataset's relatively small scale.

7. Discussion and limitations

Advantages and limitations relative to prior speckle vibrometry systems: Our system was inspired by many prior works that reduced the image domain to increase capture speed [3, 45, 53]. Closest to our work is the dual-shutter camera of Sheinin *et al.*, which can capture multiple surface points on a single container. However, since it senses only a single row, capturing multiple containers requires scan-



(a) two *empty* identical silver bottles



(b) the same vacuum flask having *different* fill levels

Figure 7. Challenging inference examples. **(a)** Two visually similar *empty* containers that exhibit different resonant frequency profiles, suggesting subtle manufacturing differences. **(b)** A different container shows nearly identical frequency responses across fill levels, indicating minimal resonance change due to thick insulation. However, noticeable differences do appear in the 100–150 Hz range, enabling inference.

ning. Conversely, our system can sense multiple containers at once, making it more practical for applications that require scanning large item sets (*e.g.*, supermarkets, assembly lines). Beyond scanning 2D grids, our system obviates the two-camera calibration of [46], avoids light loss from beam splitting, and suffers no rolling-shutter inter-frame dead times. Nevertheless, combining both approaches (*i.e.*, adding a second reference camera) can enable reducing the ROI heights to a single pixel, maximizing camera speed.

Performance analysis on various container types: Our experiments offer insights into the model's performance across container types. Firstly, we observed that highly resonant containers are *more* difficult to infer for unseen same-class containers. This is because their frequency response may be dominated by a strong resonant peak, whose frequency position can shift due to manufacturing differences (see Fig. 7(a)). For example, consider a hypothetical limiting case where two same-class containers have distinct resonant frequencies of 450 Hz and 400 Hz, respectively. In this scenario, the model would be unable to distinguish between an empty second bottle and the first containing some liquid, as liquid lowers the resonant frequency. Real containers, however, do not exhibit perfect delta-like frequency responses, making them amenable to our approach (as in Fig. 7(a)). We also observed that some vacuum flasks yield similar vibrations across fill levels due to double-wall insulation (Fig. 7(b)). Fortunately, the response differs enough at lower frequencies, allowing for a reasonable inference.

Container materials, laser safety, and audio level: We tested our method on everyday containers and captured speckle interference in most cases without modifying the packaging. However, some materials, like glass, polished metals, and ones having a very low albedo, will be less amenable to speckle-based vibrometry. For such materials, like the wine glass in Fig. 2, we placed a small sticker on the container's surface to capture the speckle. Such augmentation can be readily applied for sensing non-disposable containers (*e.g.*, in industrial factory settings). Each laser point in our prototype had 14 mW power, making direct eye exposure unsafe. [4] Thus, eye safety must be considered due to potential direct exposure from specular surfaces. We evaluated how sound volume affects vibration signal SNR (see supplementary). Results show good SNR at low volumes, suggesting the method works with minimal sound sources (*e.g.*, small compact speakers).

Generalizing to novel container classes: We present a proof-of-concept for generalizing liquid inference to unseen speaker positions, fluid levels, input sounds, and containers of the same class. However, our dataset is too small to explore broader questions: Can a model trained on enough data generalize to entirely new container classes? Classify liquid types (*e.g.*, water, soda, oil) or extend to granular materials (*e.g.*, sand)? Can vibrations serve as a 'container fingerprint' for identifying the same container across scenes? We leave these fascinating questions for future work.

8. Conclusion

We introduced a novel system that "sees" inside opaque liquid containers by combining a novel imaging system based on high-speed laser speckle vibrometry with a new deep learning architecture – the Vibration Transformer – for semantic analysis of vibration signals. We conducted an extensive experimental evaluation and provided insightful analysis of the results to validate our approach and demonstrate a proof of concept for this novel computer vision task. Our work also yielded a novel container vibration dataset that may be beneficial to the vision community beyond the scope of the current work. We hope our work inspires further research on the semantic inference of hidden scene properties, such as detecting the contents of closed packages, fruit ripeness, spoilage in sealed foods, chemical composition, and other latent attributes.

Acknowledgments We thank G. Shasha for experimental support. This work was supported by the Weizmann Center for New Scientists, the Institute of AI, and the MBZUAI-WIS Joint Program (SB).

[4] Diffuse reflections at 14 mW are generally considered safe.

References

[1] Marina Alterman, Chen Bar, Ioannis Gkioulekas, and Anat Levin. Imaging with local speckle intensity correlations: theory and practice. *ACM Transactions on Graphics (TOG)*, 40 (3):1–22, 2021. 3

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning (ICML)*, 2016. 4

[3] Nick Antipa, Patrick Oare, Emrah Bostan, Ren Ng, and Laura Waller. Video from stills: Lensless imaging with rolling shutter. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8. IEEE, 2019. 7

[4] Piyush Bagad, Makarand Tapaswi, Cees G. M. Snoek, and Andrew Zisserman. The sound of water: Inferring physical properties from pouring liquids, 2025. 2

[5] S Bianchi and E Giacomozzi. Long-range detection of acoustic vibrations by speckle tracking. *Applied optics*, 58 (28):7805–7809, 2019. 2

[6] Katherine L Bouman, Bei Xiao, Peter Battaglia, and William T Freeman. Estimating the material properties of fabric from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[7] Mingxuan Cai, Dekel Galor, Amit Pal Singh Kohli, Jacob L. Yates, and Laura Waller. Event2audio: Event-based optical vibration sensing. In *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*, 2025. 2

[8] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015. 4

[9] 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, 3

[10] Youngjun Cho, Nadia Bianchi-Berthouze, Nicolai Marquardt, and Simon J Julier. Deep thermal imaging: Proximate material type recognition in the wild through deep learning of spatial surface temperature patterns. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018. 1

[11] Coherent. Sapphire lpx 532 nm 500 mw laser. https://www.coherent.com/lasers/cw-solid-state/sapphire. Accessed: 2025-02-27. 5

[12] Aniket Dashpute, Vishwanath Saragadam, Emma Alexander, Florian Willomitzer, Aggelos Katsaggelos, Ashok Veeraraghavan, and Oliver Cossairt. Thermal spread functions (tsf): Physics-guided material classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[13] Abe Davis, Katherine L Bouman, Justin G Chen, Michael Rubinstein, Fredo Durand, and William T Freeman. Visual vibrometry: Estimating material properties from small motion in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015. 2, 3

[14] Jacob Pieter Den Hartog. *Mechanical vibrations*. Courier Corporation, 1985. 2

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (ACL)*, 2019. 4

[16] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[17] Youssef Douini, Jamal Riffi, Mohamed Adnane Mahraz, and Hamid Tairi. Solving sub-pixel image registration problems using phase correlation and lucas-kanade optical flow method. In *2017 Intelligent Systems and Computer Vision (ISCV)*, pages 1–5, 2017. 4

[18] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG). Proc. SIGGRAPH*, 37(4):1–11, 2018. 4

[19] Berthy T Feng, Alexander C Ogren, Chiara Daraio, and Katherine L Bouman. Visual vibration tomography: Estimating interior material properties from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[20] Oz Frank, Nir Schipper, Mordehay Vaturi, Gino Soldati, Andrea Smargiassi, Riccardo Inchingolo, Elena Torri, Tiziano Perrone, Federico Mento, Libertario Demi, Meirav Galun, Yonina C Eldar, and Shai Bagon. Integrating domain knowledge into deep networks for lung ultrasound with applications to COVID-19. *IEEE transactions on medical imaging*, 41(3):571–581, 2021. 4

[21] Adrián García, Víctor Toral, Álvaro Márquez, Antonio García, Encarnación Castillo, Luis Parrilla, and Diego P Morales. Non-intrusive tank-filling sensor based on sound resonance. *Electronics*, 7(12):378, 2018. 2

[22] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017. 4

[23] HOLO-OR. Diffractive beam splitters. https://holo-or.com/products/beam-splitters/. Accessed: 2025-02-27. 5

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*. pmlr, 2015. 7

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[26] Charles D Kuglin. The phase correlation image alignment method. In *IEEE International Conference on Cybernetics and Society*, 1975. 4

[27] Haejoon Lee and Aswin C Sankaranarayanan. Spectral subsurface scattering for material classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1

[28] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[29] Peter F Lichtenwalner and Donald A Sofge. Local-area damage detection in composite structures using piezoelectric transducers. In *Smart Structures and Materials 1998: Industrial and Commercial Applications of Smart Structures Technologies*, pages 509–515. SPIE, 1998. 3

[30] Chao Liu and Jinwei Gu. Discriminative illumination: Per-pixel classification of raw materials based on optimal projections of spectral brdf. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(1), 2013. 1

[31] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, pages 674–679, 1981. 4

[32] Mikrotron. Eosens2.0mcx12-cm machine vision camera. https://mikrotron.de/en/high-speed-cameras/mik-camera-detail.php?id=EoSens2.0MCX12-CM. Accessed: 2025-02-27. 5

[33] Sriram Narayanan, Mani Ramanagopal, Mark Sheinin, Aswin C Sankaranarayanan, and Srinivasa G Narasimhan. Shape from heat conduction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1

[34] Mark of the Unicorn (MOTU). Ultralite-mk5 sound card. https://motu.com/en-us/products/gen5/ultralite-mk5/. Accessed: 2025-02-27. 5

[35] Alexander C. Ogren, Berthy T. Feng, Jihoon Ahn, Katherine L. Bouman, and Chiara Daraio. Visual surface wave elastography: Revealing subsurface physical properties via visible surface waves. *arXiv preprint arXiv:2507.09207*, 2025. 2

[36] JJ Pearson, DC Hines Jr, S Golosman, and CD Kuglin. Video-rate image correlation processor. In *Applications of digital image processing*, pages 197–205. SPIE, 1977. 4

[37] Mani Ramanagopal, Sriram Narayanan, Aswin C Sankaranarayanan, and Srinivasa G Narasimhan. A theory of joint light and heat transport for lambertian scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[38] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *IEEE spoken language technology workshop (SLT)*, 2018. 4

[39] Steve Rothberg, JR Baker, and Neil A Halliwell. Laser vibrometry: pseudo-vibrations. 1989. 3

[40] Subhankar Roy, Willi Menapace, Sebastiaan Oei, Ben Luijten, Enrico Fini, Cristiano Saltori, Iris Huijben, Nishith Chennakeshava, Federico Mento, Alessandro Sentelli, Emanuele Peschiera, Riccardo Trevisan, Giovanni Maschietto, Elena Torri, Riccardo Inchingolo, Andrea Smargiassi, Gino Soldati, Paolo Rota, Andrea Passerini, Ruud J. G. van Sloun, Elisa Ricci, and Libertario Demi. Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. *IEEE transactions on medical imaging*, 39(8):2676–2687, 2020. 4

[41] Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, and Chandra Kambhamettu. Material classification with thermal imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4649–4656, 2015. 1

[42] Vishwanath Saragadam and Aswin C Sankaranarayanan. Krism—krylov subspace-based optical computing of hyperspectral images. *ACM Transactions on Graphics (TOG)*, 38 (5):1–14, 2019. 1

[43] Vishwanath Saragadam and Aswin C Sankaranarayanan. Programmable spectrometry: Per-pixel material classification using learned spectral filters. In *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*, 2020.

[44] Vishwanath Saragadam, Michael DeZeeuw, Richard G Baraniuk, Ashok Veeraraghavan, and Aswin C Sankaranarayanan. Sassi—super-pixelated adaptive spatio-spectral imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 43(7):2233–2244, 2021. 1

[45] Mark Sheinin, Dinesh N. Reddy, Matthew O'Toole, and Srinivasa G. Narasimhan. Diffraction line imaging. In *European Conference on Computer Vision (ECCV)*, pages 1–16, 2020. 7

[46] Mark Sheinin, Dorian Chan, Matthew O'Toole, and Srinivasa G. Narasimhan. Dual-shutter optical vibration sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 8

[47] Brandon M Smith, Pratham Desai, Vishal Agarwal, and Mohit Gupta. Colux: Multi-object 3d micro-motion analysis using speckle imaging. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 3

[48] Creative Technology. Pebble v2 speakers pair. https://us.creative.com/p/speakers/creative-pebble-v2. Accessed: 2025-02-27. 5

[49] Thorlabs. Anamorphic prism pairs. https://www.thorlabs.com/thorproduct.cfm?partnumber=PS873-A. Accessed: 2025-02-27. 5

[50] Shoji Tominaga and Tetsuya Yamamoto. Metal-dielectric object classification by polarization degree map. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, 2008. 1

[51] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop (SSW)*, 2016. 4

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 4

[53] Gil Weinberg and Ori Katz. 100,000 frames-per-second compressive imaging with a conventional rolling-shutter

camera by random point-spread-function engineering. *Optics Express*, 28(21):30616–30625, 2020. 7

[54] Justin Wilson, Auston Sterling, and Ming C. Lin. Analyzing liquid pouring sequences via audio-visual neural networks. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7702–7709, 2019. 2

[55] Nan Wu and Shinichiro Haruyama. Fast motion estimation of one-dimensional laser speckle image and its application on real-time audio signal acquisition. In *2020 the 6th International Conference on Communication and Information Processing*, pages 128–134, 2020. 2

[56] 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.

[57] Zeev Zalevsky, Yevgeny Beiderman, Israel Margalit, Shimshon Gingold, Mina Teicher, Vicente Mico, and Javier Garcia. Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern. *Optics express*, 17(24):21566–21580, 2009. 2, 3

[58] Tianyuan Zhang, Mark Sheinin, Dorian Chan, Mark Rau, Matthew O'Toole, and Srinivasa G. Narasimhan. Analyzing physical impacts using transient surface wave imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2