

Revealing Subtle Heat Flows All Around Us Using Microbolometric Videos

Mani Ramanagopal, Akihiko Oharazawa, Sriram Narayanan, Zeqing Yuan, and Srinivasa G. Narasimhan

Abstract—Thermal videos captured by low-cost microbolometric cameras are highly noisy. Additionally, in the absence of strong heat sources, the scenes have low thermal contrast making heat transport analysis even harder. In this paper, we demonstrate that by properly modeling the bolometric pixel behavior, it is possible to overcome both the high noise and low contrast in thermal signals to reveal subtle heat flows in everyday scenes from noisy measurements. Our key insight is that microbolometers always return temporally correlated measurements and each pixel's response function varies across the focal plane array and with operating conditions. This means that any spatial processing, perhaps to extract features or estimate motion or align patches across frames, works against an artificially elevated noise floor established by this spatial non-uniformity in pixel responses. Our simple solution is to perform a per-pixel temporal 1D wavelet transform and then apply spatial processing to the resulting coefficient maps. The clean video is reconstructed from the processed coefficient maps using the canonical dual frame. This approach is highly parallel and achieves a spatially coherent video even in the presence of strong pixel-, row-, and column-wise drifts, as well as additive and multiplicative noise. Thanks to our denoising, we use simple operations such as time derivatives and spatial iso-contours of pixel intensity fields to reveal subtle heat flows that are almost imperceptible in the captured videos. Beyond visualizing heat transport, we show improvement on downstream tasks such as shape from heat conduction, visible-thermal intrinsic image decomposition and separation of thermal emission and reflection. Our denoising can also broaden the applicability of computational imaging approaches that use optical filtering (polarization, spectral filtering, beam splitters, etc.) in the thermal spectrum which reduce SNR of captured data even further. We hope our approach will help move inverse heat transport research out of controlled and specialized industrial domains to complex everyday scenes all around us.

Index Terms—Uncooled Microbolometers, Non-Uniformity Correction, Thermal Computational Imaging

1 INTRODUCTION

HEAT transport is happening all around us, all the time, even in a dark room or a cold freezer. Even subtle heat flows, such as due to a gentle breeze (Fig. 1), should theoretically encode rich information about the scene. Despite their omnipresence, thermography, or the study of heat transport, is primarily used in specialized applications such as defense and industrial inspection where the targets have high temperature or medical imaging and remote sensing where controlled setups and expensive cooled thermal cameras can be afforded. It is encouraging that recent works have used low-cost thermal cameras to recover shape [1], [2], [3], material [4], and lighting information [5], [6], as well as decompose thermal light transport [7], [8]. However, these methods still rely on excessive heating, controlled laboratory setups or strong priors.

These low-cost thermal cameras are uncooled microbolometers and they have a reputation for being noisy. But the Noise Equivalent Temperature Difference (NETD), the primary figure of merit for thermal camera hardware, is as low as 20 – 50 mK for modern microbolometers. In other words, current thermal camera hardware is much more sensitive than the algorithms we use with them lead us to believe. In this paper, we show that proper modeling of the bolometric pixel behavior can allow simple algorithms to reveal subtle heat flows all around us, especially when visualized using an appropriate representation.

Microbolometers differ fundamentally from CMOS sensors. Each pixel is a temperature-dependent resistor that heats and cools continuously in response to incident flux. They are constantly exposed to the scene, have no reset at the end of readout and their measurements are temporally correlated due to thermal inertia. As the sensor is not temperature-stabilized, the response of the bolometric pixels drifts slowly and independently from their neighbors. Similarly, the readout electronics which are shared across rows and columns drift as well, which introduces axis-correlated non-uniformities. It is these diverging temporal drifts that manifest spatially as Fixed Pattern Noise (FPN). In this paper, we embrace this inherent temporal nature of microbolometers rather than rely on ever stronger and computationally expensive spatial priors.

Our key insight is to first treat each pixel in a thermal video as a separate 1D array and apply the Dual-Tree Complex Wavelet Transform (DT-CWT) [9] to it. Note that a pixel's temporal drift can be well approximated by a low order polynomial. As DT-CWT has multiple vanishing moments (depending on the choice of filters), any pixel-specific drift and low frequency signal gets accumulated into the approximation coefficients [10] as we decompose to larger depths. The approximation coefficients at the coarsest level have a higher SNR, which enables spatial denoisers to better separate drift from signal. This separation of temporal dynamics in a thermal video across multiple scales, afforded by the wavelet transform, allows us to simplify denoising long videos into spatially smoothing a sparse set of coefficient maps. For spatial processing of coefficient maps, we propose a two stage approach. First, for low SNR scenes, we

- All authors are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 15213.
E-mail: {manikans,aoharaza,nnsriram,leoyuan,srinivas}@andrew.cmu.edu
- Project webpage: <https://revealing-subtle-heat.github.io/>

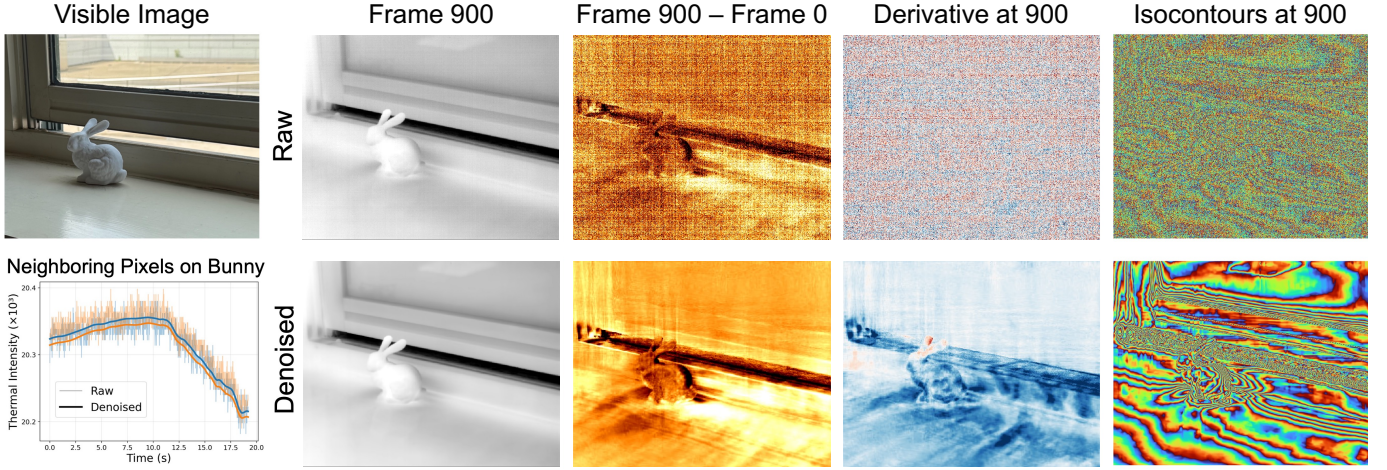


Fig. 1: Heat transport occurs all around us all the time. A gentle draft of wind blowing through the window opening cools the sill, bunny and the table. The median value of the temporal standard deviation of the above video is < 200 mK. Bolometric thermal cameras are highly noisy and low contrast, so such subtle heat transients are thought to not be captured. Our work shows that proper modeling of the bolometric pixel behavior leads to a simple wavelet domain algorithm to denoise such videos revealing subtle heat transients that were there all along. While the denoised frame does not seem too different from the captured frame, common operations necessary for heat transport modeling and analysis including heat accumulated over time (Frame 900 minus Frame 0), the time derivative of temperature and the evolution of temperature iso-contours clearly reveal heat flows. See supplementary video for continuous flows.

develop a custom destriping process. Then, we use a non-local method, such as BM3D [11], as a gaussian denoiser. While any denoiser (including deep learning based) could be applied here, BM3D proved effective in our experiments in recovering subtle spatial structures in the coefficient maps.

Beyond noise, a second criticism of microbolometer videos is low image contrast in the absence of strong heat sources in the scene. But thanks to effective denoising, we show that simple temporal operations and visualizations can reveal even the most subtle heat flows. These operations include the cumulative intensity change since the start of the video (current frame minus first frame), temporal derivative of pixel intensity, or spatial iso-contours of intensity fields. These quantities are not just helpful in visualization but are crucial to modeling and analyzing heat transport.

We demonstrate the effectiveness of our Temporal-Spatial Video Processing (TSVP) framework using microbolometer videos of everyday scenes to reveal often surprising subtle heat flows even when the noise in the captured videos seem to be overwhelming. Consider the *bunny draft* scene in Fig. 1 where a gentle wind draft slightly cools the window sill, object and table. The temporal visualizations of cumulative intensity change, temporal derivative and iso-contours after denoising show the heat transport even though the corresponding visualizations of the captured video look somewhat hopeless. We illustrate our approach in a similar way in diverse indoor and outdoor scenes in Sections 5 and 6 to reveal signals such as vehicle exhaust, sharp text on books, thermal reflections from glossy surfaces even when the signals seem to be buried in noise.

We also demonstrate our approach when the thermal cameras are optically filtered (using beam splitters, polarizers, spectral filters, etc.), with even lower SNR, better

enabling computational imaging methods that use such filtering. Finally, we show the benefit of our algorithm for downstream tasks such as shape from heat conduction, separation of thermal reflection and emission, and visible-thermal intrinsic image decomposition. We hope this work spurs research on general inverse heat transport in everyday complex scenes with low-cost microbolometric cameras.

2 RELATED WORK

2.1 Thermal Computational Imaging

Thermal imaging has attracted growing interest in computer vision because it captures wavelengths beyond the visible spectrum and thereby grants access to physical quantities — material properties, temperature distributions, and radiative characteristics — that are otherwise hidden. For instance, JoLHT-Video [5] formalizes the coupling between a visible image of a scene and its thermal video. Several works exploit thermal-specific physical cues, including intrinsic thermal emission [3], polarization arising from thermal emission and reflection [12], [13], and heat conduction for object reconstruction that is independent of reflectance and scene lighting [1], resolving long-standing ambiguities in physics-based vision [2]. These ideas have been extended to material classification [4], intrinsic image decomposition [5], [6], pose estimation [14], tracking [15], and the separation of reflected and emitted thermal radiation [7], [8], demonstrating the breadth of information that can be recovered from the thermal domain. However, uncooled microbolometers — the only low-cost option for long-wave infrared (LWIR) capture — introduce sensor-specific noise that significantly degrades image quality and limits the practical reach of these methods.

2.2 Scene-based and Shutter-based NUC

Non-Uniformity Correction (NUC) is a fundamental preprocessing step that compensates for pixel-wise variations in sensor response, and is typically classified as either shutter-based or scene-based. Shutter-based methods, most notably Flat Field Correction (FFC), estimate per-pixel offset by periodically capturing reference frames with a mechanical shutter closed over the sensor. While effective, FFC interrupts normal imaging and is therefore unsuitable for continuous capture, particularly when the goal is to observe slow thermal phenomena. Scene-based methods, in contrast, estimate correction parameters directly from observed image sequences without dedicated calibration frames [16]. Early work by Hardie et al. [17], [18] leverages motion and frame registration across video sequences to estimate pixel-wise parameters, but these approaches assume sufficient scene motion and accurate alignment — assumptions that break down in the static, low-contrast scenes we target.

2.3 Thermal Image and Video Restoration

Noise removal in uncooled thermal imaging has been addressed both as single-image NUC and as multi-frame video restoration. Single-image approaches include statistical methods [19] and CNN-based methods [20]; because they operate on one observation, they struggle to disentangle structured noise from scene content. Multi-frame methods exploit temporal consistency across frames to achieve stronger separation [21], [22], [23]. A complementary line jointly removes random and fixed-pattern noise through spatiotemporal video filtering [24], and general-purpose deep video denoisers adopt a similar spatiotemporal philosophy [25]. DeepIR, for example, assumes small camera motion so that a clean latent image can be optimized jointly with per-pixel noise parameters. MDIVDnet [26] combines spatio-temporal features with a physics-based noise generator for end-to-end restoration. Recent progress has been dominated by deep networks, including CNN-based designs [20], [27], [28], [29], [30], adversarial learning [31], and diffusion models [32]. Supervised training, however, is constrained by the scarcity of clean thermal ground truth. Physics-based noise synthesis has been proposed as a remedy [21], [26], yet its success hinges on the accuracy of the underlying noise model and its ability to capture the non-stationary drift processes we observe in real data (Sec. 3).

2.4 Wavelet-based Thermal Image Restoration

Wavelet-based processing has long been used for thermal data. However, most prior methods apply wavelet transforms along the spatial dimensions, to identify stripe noise and suppress gaussian noise. In contrast, we empirically demonstrate the surprising utility of first applying temporal wavelet transform before any spatial processing. These prior works could be used to improve the spatial processing applied to the temporal coefficient maps in our framework.

Prior work has explored wavelet-domain denoising with complex wavelet transforms and optimization-based refinements [33], as well as statistical models such as the Gaussian Scale Mixture [34]. Wavelet representations have also been applied to NUC by isolating sensor-specific noise in the frequency domain [35]. More recent work integrates wavelets

with deep learning, combining joint spatial-wavelet features and attention mechanisms to improve the separation of noise from detail [36], [37]. These methods often demonstrate their performance on high SNR scenes. In contrast, our Temporal-Spatial Video Processing reveals subtle heat flows in extremely noisy thermal videos.

3 TEMPORAL DYNAMICS OF A MICROBOLOMETER

A precise forward model is hard to pin down for microbolometers due to the fundamental stochastic nature of heat transport within them and the many sources of aberrations that vary with operating conditions. We therefore take an empirical approach, analyzing the raw output from a real microbolometer to motivate our algorithmic choices. Our analysis reveals two key facts: (1) commonly used noise models do not model drifts, and (2) low-frequency temporal drift is in fact a major contributor to high frequency spatial aberrations.

3.1 Empirical Analysis of Raw Microbolometer data

We captured 1024 frames at 60 Hz of a plain wall at thermal equilibrium using a FLIR Boson+ thermal camera (640×512 resolution, NETD < 20 mK). The recording began several minutes after power-on; an FFC was performed immediately before capture and then disabled for its duration. We also disabled the on-board spatio-temporal filters including Temporal Filter and the Supplemental Column/Row Noise Reduction filters to access the raw sensor output while retaining all fixed factory calibrated non-uniformity corrections.

Let $I[n, i, j]$ denote the pixel intensity at row i , column j and frame n . To analyze axis-correlated temporal dynamics, we define *row-mean signals* $I_r[n, i, \cdot] = \frac{1}{W} \sum_{j=1}^W I[n, i, j]$ and *column-mean signals* $I_c[n, \cdot, j] = \frac{1}{H} \sum_{i=1}^H I[n, i, j]$. We use the autocorrelation function (ACF) to identify temporal trends in these signals. Specifically, the ACF between $x_1[n]$ and $x_2[n]$ at lag τ is defined as

$$\text{ACF}_{x_1, x_2}(\tau) = \frac{1}{N - \tau} \sum_{n=1}^{N-\tau} \frac{(x_1[n] - \mu_1)(x_2[n + \tau] - \mu_2)}{\sigma_1 \sigma_2}, \quad (1)$$

where μ_k and σ_k are the temporal mean and standard deviation of $x_k[n]$.

As the scene itself is static and in thermal equilibrium, the thermal intensity at every pixel must be constant in time. Figure 2 shows pixel-intensity plots and ACF curves for two neighboring pixels, two neighboring rows and two neighboring columns. The two horizontal dashed lines denote the 95% confidence interval (CI) for white noise $\pm 1.96/\sqrt{N}$. At pixel, row and column levels, the ACF lies outside this band, confirming statistically significant low frequency trends. One might expect this to be the scene equilibrium slowly drifting. However, we notice that these drifts differ within each pair. Since heat conduction would smooth out temperature differences between neighboring points on a wall, the drifts in this case must be due to the temporal dynamics of the microbolometer itself. Such divergence is especially prominent at the column level.

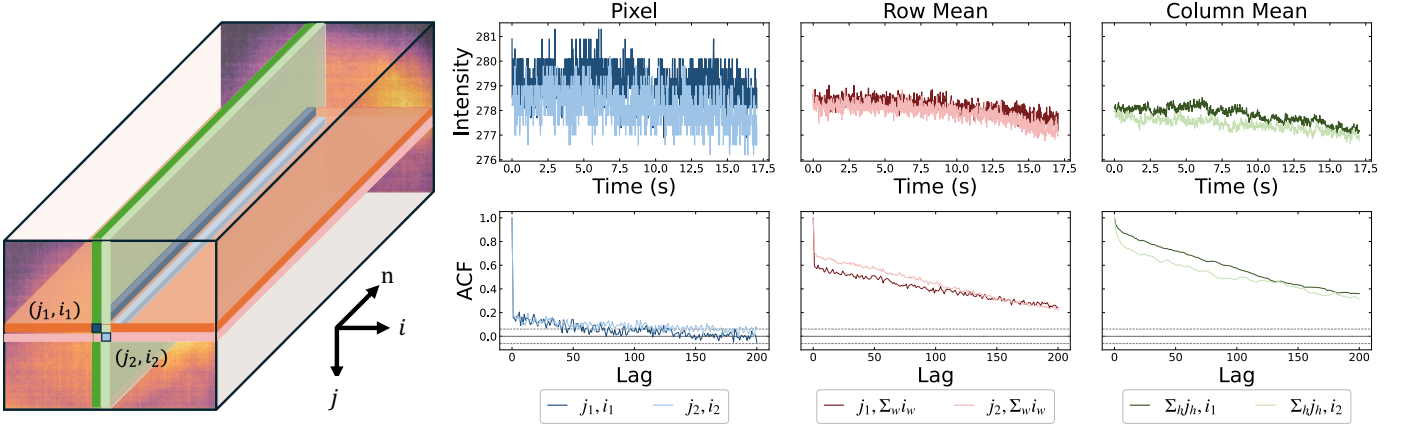


Fig. 2: (left) An illustration of the temporal dynamics of raw data from a real microbolometer observing a plain wall at thermal equilibrium. (right) The top row shows pixel intensities of two diagonally neighboring pixels, mean intensities of their corresponding rows and that of their columns. The bottom row shows their corresponding autocorrelation function (ACF) which reveals the *diverging* low frequency trends between neighboring pixels / rows / columns present in real thermal cameras even in ambient indoor conditions.

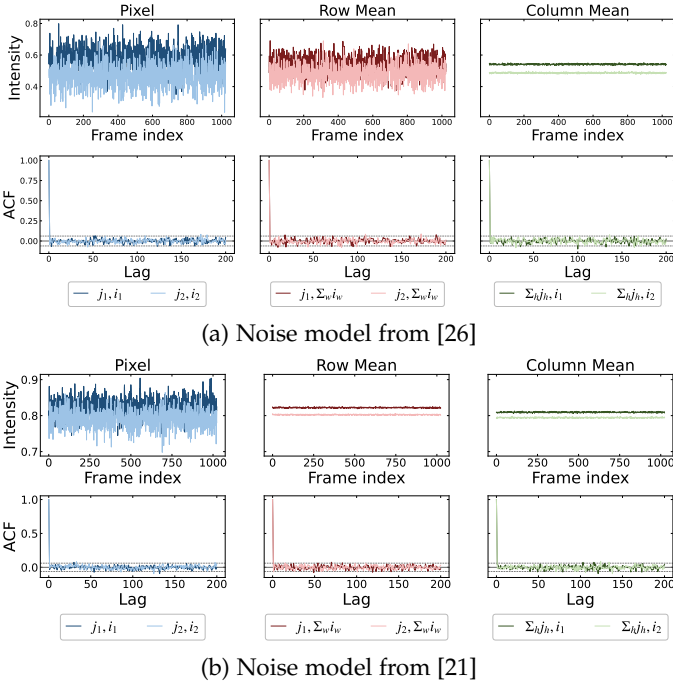


Fig. 3: Temporal intensity plots and their corresponding autocorrelation function (ACF) plots for two commonly used noise models for thermal cameras in computer vision. These do not capture the drift in temporal dynamics of real microbolometers as seen in Fig. 2.

3.2 Insufficiency of Current Noise Models

There are broadly two classes of noise models used in thermal computer vision. To test these noise models, we consider a scene at perfect equilibrium with all ground truth pixel values at $I^* = 0.5$. The **additive model** [26], [38], [39] decomposes the video as

$$I[n, i, j] = I^*[n, i, j] + N_g[n, i, j] + N_l[n, i, \cdot] + N_c[\cdot, j] + N_r[i, \cdot] + N_h[i, j], \quad (2)$$

where I^* is the true scene signal, N_g is additive white Gaussian noise, N_l is temporal line noise, N_c and N_r are fixed spatial stripe noises, and N_h is hill noise [26]. The **affine model** [21], [40] accounts for per-pixel gain and offset as:

$$I[n, i, j] = (G[i, j]I^*[n, i, j] + O[i, j]) + N_e[n, i, j], \quad (3)$$

where G and O are the pixel-wise gain and offset relating input radiant flux to output voltage and N_e is the electronic readout noise.

The pixel intensity and ACF plots for both these noise models are shown in Fig. 3. As both these models use only Gaussian processes even in the time dimension, the ACF plots at the pixel, row and column level all are consistently within the 95% CI bands. This is qualitatively inconsistent with what we observed in real data (Fig. 2). Algorithms trained or designed under such models would therefore struggle to perform well on real microbolometers, especially at low SNR.

3.3 Temporal Drift and the Fixed Pattern Noise

In visible cameras, the response of every pixel to incident flux is nearly identical. As a result, the difference in intensity between neighboring pixels exposed to identical scene flux is primarily the read noise which is zero-mean Gaussian, although with a higher variance. This is not the case for microbolometers, where neighboring pixels experience diverging drifts. Therefore, spatial differences in a thermal video of an ideal uniform thermal target would have non-zero expected value equal to the difference in expected value of each pixel. As these drifts are low frequency in time, they appear to be “fixed” when viewing a dynamic scene, and hence the name Fixed Pattern Noise (FPN).

As these drifts cannot be eliminated at the hardware-level, all microbolometers rely on either a shutter-based or scene-based NUC to mitigate it. All shutter-based Flat Field Correction (FFC) techniques introduce sharp jumps in pixel statistics before and after the FFC, which becomes particularly problematic when studying slow thermal phenomenon

of low contrast scenes. On the other hand, scene-based NUC results in a burn-in effect where faint edges from static parts of the scene stick with the video. These algorithms alter the pixel statistics even before they are recorded by the camera. In this work, we instead develop algorithms that operate directly on the raw thermal video feed from a microbolometer.

3.4 Spatial Coherence of Temporal Signals

The core issue with applying spatial denoisers to raw thermal images is that the response of any spatial kernel includes the response from both the signal and the drift. Our simple solution is to first perform a temporal operation that projects the signal, drift and noise into a common set of functions, before doing any spatial processing. Specifically, we apply 1D wavelet transform along time at each pixel independently. The slow pixel-specific drifts, along with low frequency signal, get accumulated to their respective approximation coefficients, leaving the detail coefficients at finer scales robust to temporal drifts. As the approximation coefficients have a higher signal to noise ratio, spatial denoising algorithms are able to better separate drift from the signal in the coefficient domain.

Note that, we do not aim to recover absolute radiometric intensities in the recovered video. Instead, we seek thermal videos that remain spatially coherent at *all times and all temporal scales*. This is particularly applicable for real microbolometers as their pixel intensities are only related to the current pixel temperature. The finite heat capacity of the pixel implies that its temperature cannot change instantaneously even if the scene changes instantaneously due to a hot object moving across a pixel.

4 TEMPORAL-SPATIAL VIDEO PROCESSING

Our framework, which we call Temporal-Spatial Video Processing (TSVP), chains a temporal transform with a spatial denoiser applied to each resulting coefficient map. The design exploits the unique temporal structure of microbolometers (Sec. 3) while leveraging the extensive literature on 2D image denoising. An overview is shown in Fig. 4 and the algorithm is summarized in Alg. 1.

4.1 Temporal Transform with Vanishing Moments

Let $\mathbf{I}_{pre} \in \mathbb{R}^{N \times H \times W}$ denote the preprocessed thermal video (Sec. 4.4). The intensities at each pixel is a vector of N values that is a function of the true signal from the scene, drift in response at the pixel / row / column levels, and noise at all three levels - all mixed together.

Desired transform properties. The primary job of the temporal transform is to map these N numbers into M coefficients such that spatial processing at each of the M coefficient indices preserves the signal. To do this efficiently requires:

- *Shift invariance:* neighboring pixels witness an event / edge at slightly different times. To preserve the timing of events, the transform should produce coefficients that vary smoothly for shifted inputs.

Algorithm 1 Temporal Spatial Video Processing

```

1: procedure TSVP( $\mathbf{I}$ )
2:    $\mathbf{I}_{pre} \leftarrow \text{preTransform}(\mathbf{I})$  ▷ Sec. 4.4
3:    $(\mathbf{A}_s, \mathbf{D}_\ell) \leftarrow \text{forwardTransform}(\mathbf{I}_{pre})$  ▷ Sec. 4.1
4:   // Process Approximation Coefficients ▷ Sec. 4.3.1
5:    $\forall m, \hat{\mathbf{A}}_s^1[m] \leftarrow \text{spatialProcess}(\mathbf{A}_s[m])$ 
6:    $\bar{N}_s \leftarrow \sum_{m=1}^{M_s} (\hat{\mathbf{A}}_s^1[m] - \mathbf{A}_s[m])$ 
7:    $\forall m, \hat{\mathbf{A}}_s[m] \leftarrow \text{spatialProcess}(\mathbf{A}_s[m] - \bar{N}_s)$ 
8:   // Process Detail Coefficients ▷ Sec. 4.3.2
9:    $\forall m, \ell, \hat{\mathbf{D}}_\ell[m] \leftarrow \text{spatialProcess}(\mathbf{D}_\ell[m])$ 
10:   $\forall m, \ell, \hat{\mathbf{D}}_\ell[m] \leftarrow \text{softShrink}(\hat{\mathbf{D}}_\ell[m])$ 
11:   $\hat{\mathbf{I}}_{post} \leftarrow \text{inverseTransform}(\hat{\mathbf{A}}_s, \hat{\mathbf{D}}_\ell)$ 
12:   $\hat{\mathbf{I}} \leftarrow \text{postTransform}(\hat{\mathbf{I}}_{post})$  ▷ Sec. 4.4
13:  return  $\hat{\mathbf{I}}$ 
14: end procedure
15: procedure SPATIALPROCESS( $X$ )
16:   $\tilde{X} \leftarrow \text{destripeImage}(X)$  ▷ Sec. 4.3.3
17:   $\tilde{X} \leftarrow \text{denoiseImage}(\tilde{X})$  ▷ BM3D [11]
18:  return  $\tilde{X}$ 
19: end procedure

```

- *Linear phase:* spatial denoisers rely on coefficient magnitude to differentiate signal from noise. The transform should produce larger coefficients near time instances when a transient happens.
- *Limited redundancy:* ideally the transform is orthogonal (i.e. $M = N$). If not, lower redundancy results in fewer coefficients to process.
- *Compact support:* finite-length filters enable efficient computation.

The standard critically sampled Discrete Wavelet Transform (DWT) is not shift-invariant due to the downsampling operation at every scale, while the Stationary Wavelet Transform (SWT) is highly redundant ($M \gg N$). We therefore adopt the **Dual-Tree Complex Wavelet Transform (DT-CWT)** [9] from the signal processing literature as an engineered compromise: it is nearly shift-invariant, has near-linear phase and compact support with only $2 \times$ redundancy irrespective of the decomposition depth.

Mathematical formulation. Applying a 1D DT-CWT along the temporal dimension at each spatial location independently yields an approximation coefficient tensor $\mathbf{A}_s \in \mathbb{R}^{M_s \times H \times W}$ and a set of detail coefficient tensors $\{\mathbf{D}_\ell \in \mathbb{C}^{M_\ell \times H \times W}\}_{\ell=1}^s$, where s is the maximum decomposition depth, M_s is the number of approximation coefficients and M_ℓ is the number of detail coefficients at each scale. The video is thus expanded as

$$I[n, i, j] = \sum_{m=1}^{M_s} \mathbf{A}_s[m, i, j] \varphi_{s,m}[n] + \sum_{\ell=1}^s \sum_{m=1}^{M_\ell} \mathbf{D}_\ell[m, i, j] \psi_{\ell,m}[n], \quad (4)$$

where $\varphi_{s,m}$ and $\psi_{\ell,m}$ are the set of scaling functions and wavelet functions that are shared across all spatial locations. In all our experiments we use the default implementation from the `dtcwt` Python module [41] with Near-Symmetric 5,7-tap filters for level 1 and Kingsbury's 10,10-tap QShift filters at levels ≥ 2 .

Role of vanishing moments. A wavelet with K vanish-

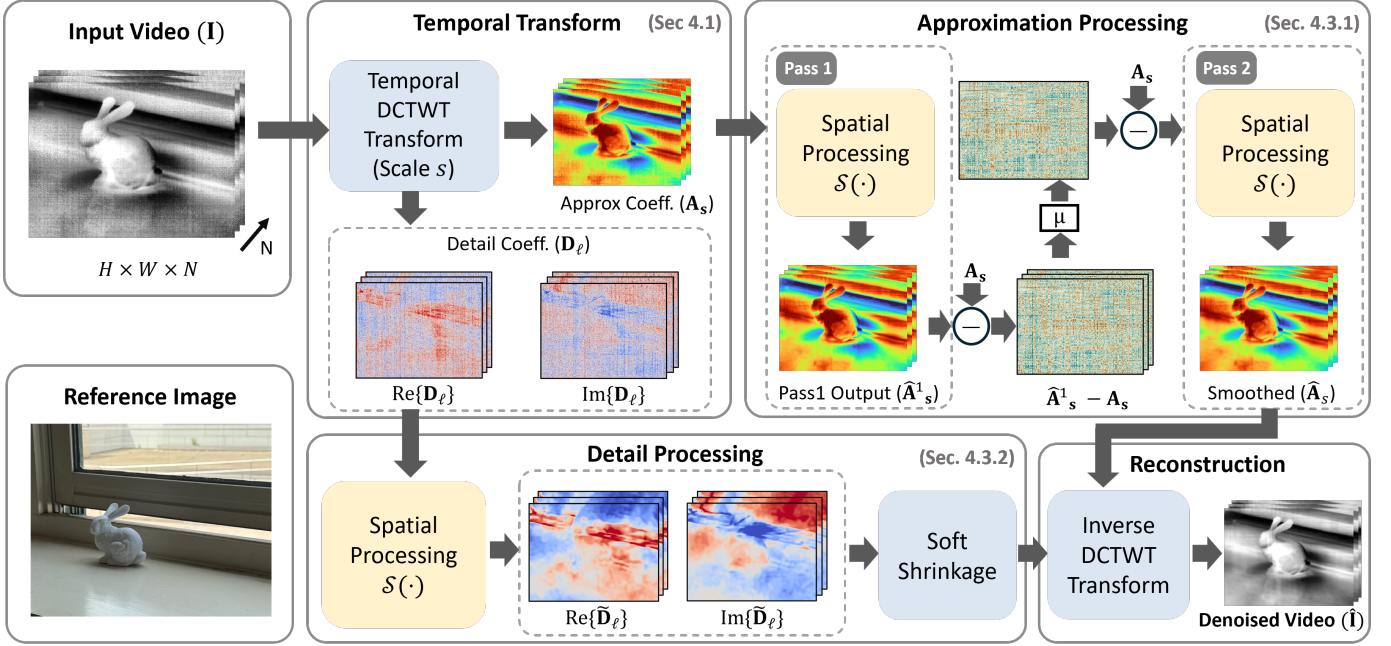


Fig. 4: Overview of our Temporal-Spatial Video Processing framework. A 1D temporal transform is applied to the video along the time dimension. The resulting coefficient tensors are spatially processed (destriped and denoised) as individual 2D slices. The non-uniformities due to per-pixel, per-row, and per-column aberrations in the microbolometer get distributed over all temporal coefficients while the signal gets concentrated to a few sparse coefficients. To separate the fixed pattern noise from the low frequency content of the signal, we adopt a two pass strategy. The mean residual from the first pass corresponds to the fixed pattern noise which is subtracted from the approximation coefficients before the second pass. For low SNR scenes, we developed a custom destriping algorithm that complements an off-the-shelf Gaussian denoiser.

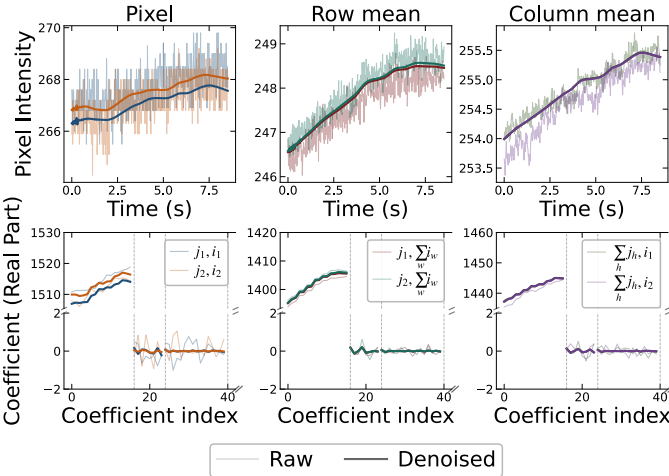


Fig. 5: Time domain (top) and wavelet domain plots (bottom) of two diagonally adjacent pixels, their corresponding row-means and column-means from the *bunny draft* scene with first 512 frames. Quantities from the noisy input video are shown in light colors while those from the denoised video are shown in dark solid lines. The pixel depicted in blue appears to be above the orange pixel in the noisy video. But this pixel had in fact drifted farther with respect to its neighbors. In the denoised video, the blue pixel is in fact lower than the orange pixel. Note that the detail coefficients of the denoised video are significantly smaller than the noisy coefficients. The row-mean and column-mean plots snap onto each other after denoising.

ing moments satisfies

$$\int t^p \psi(t) dt = 0, \quad p = 0, \dots, K-1. \quad (5)$$

This holds at every decomposition level in a Multiresolution Analysis (MRA). Therefore, slow variations in intensity that can be well represented by a low-order polynomial produce zero detail coefficients. With large s , such slow trends get accumulated into the few approximation coefficients $\mathbf{A}_s$ at the coarsest scale while the detail coefficients $\mathbf{D}_\ell$ isolate the non-polynomial and sudden variations.

For natural scenes, the incident flux in continuous time is well representable using piecewise smooth functions. Specific to microbolometers, the pixel intensities are only a function of pixel temperature which follows the incident flux with a lag due to its thermal inertia. As pixels have finite heat capacity, the pixel temperature is even more likely to be a piecewise smooth continuous function. For such functions, the vanishing moments property of wavelets results in only a sparse set of detail coefficients being above the noise floor. So for computational efficiency, we can skip coefficient maps whose Frobenius norm at quarter spatial resolution is below a threshold (tuned once per camera model) and treat them as zeros. The denoised video is reconstructed using the canonical dual frame of DT-CWT with the spatially processed coefficient tensors.

4.2 Signal, Drift and Noise in the Coefficient Domain

The coefficient tensors contain all the information present in the noisy video $\mathbf{I}_{pre}$ whether it is signal, drift or noise.

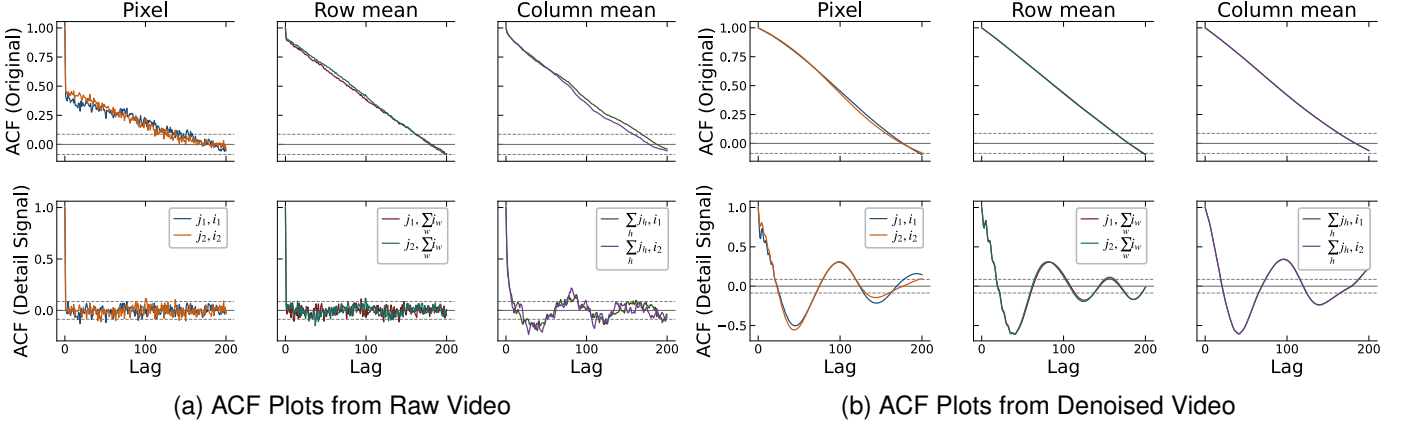


Fig. 6: Autocorrelation functions of the original (top) and detail signals (bottom) for neighboring pixels / rows / columns computed from the noisy (left) and denoised videos (right). The ACF of the detail signal at both pixels of the noisy video is statistically similar to Gaussian noise, while that of the denoised signal shows clear low frequency trends. The ACF plots of the row means and column means are diverging in the noisy video but overlap each other in the denoised video, implying that our algorithm removes spatially varying drift while recovering detail signal that was buried under the noise floor.

While additive Gaussian noise has a flat spectrum that is easier to distinguish from signal based on coefficient magnitude, drift has a $1/f$ spectrum with more energy at lower frequencies, just like natural signals. This also means that standard wavelet-domain denoising methods, which use some variant of a thresholding operator on detail coefficients alone, is not suitable for removing drift. The core issue is that a single pixel’s coefficients does not have enough information to distinguish between signal and drift.

We instead look spatially at the 2D coefficient maps at each temporal index. Our key assumption is that *true signal is spatially coherent* while *drift is spatially incoherent* in the coefficient domain. And this holds for every coefficient slice in all $\mathbf{A}_s$ and $\mathbf{D}_\ell$. This opens the door to using the vast literature on 2D image denoising to process the coefficient maps. To illustrate its effectiveness, we first discuss its performance on the *bunny draft* scene and then describe our specific spatial processing algorithms in Sec. 4.3.

Consider two diagonally neighboring pixels on the face of the bunny. The top left plot in Fig. 5 shows the noisy intensities in light color and it appears the blue pixel is above the orange pixel. In the bottom left plot, we see the two light colored plots are more clearly separated in the approximation coefficients, and the blue plot is above the orange plot. However, after denoising based on spatial coherence, the thick blue plot is below the orange one. It so happens that the blue pixel had drifted much further than its neighboring pixels, which appears as salt-and-pepper noise in the coefficient maps (see Fig. 7), and spatial denoisers are able to correct for it. At the row and column levels, spatial denoising in the coefficient domain is able to pull their temporal plots close together.

Next, we use ACF plots to analyze the temporal dynamics at the pixel / row / column levels. The top row in Fig. 6 shows these plots for the noisy (left) and denoised (right) signals. We see more slowly but jointly decaying correlations with lag in the denoised videos as is expected for heat transport phenomenon. In the bottom row, we analyze the temporal dynamics of the corresponding detail

signals, which is obtained by synthesizing a signal from only the detail coefficients. At the pixel and row levels, the ACF values fall completely within the 95% CI bands, meaning these signals look like pure noise. However, upon denoising, the correlations between neighboring pairs show real signal at pixel / row / column levels. These observations further emphasize the role of spatial processing in extracting signal from coefficients that look like noise, which we discuss next.

4.3 Spatial Processing of 2D Coefficient maps

Let $\mathcal{S}(\cdot) : \mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{H \times W}$ be a generic spatial denoising function. We will first explain how we use $\mathcal{S}$ to denoise the coefficient tensors and then provide details about our specific choice.

4.3.1 Approximation Coefficients

A constant offset in the pixel’s response causes an offset to all of its approximation coefficients. In low-SNR scenes, the temporal variation in the signal could in fact be smaller than the relative offset of the pixel with its neighbor. But we have no apriori knowledge of what is drift vs signal in the approximation tensor $\mathbf{A}_s$. We therefore adopt a two-pass strategy.

In the first pass, we compute $\hat{\mathbf{A}}_s^1[m] = \mathcal{S}(\mathbf{A}_s[m])$ for each m . The mean residual across all approximation maps, $\bar{N}_s = \frac{1}{M_s} \sum_m (\hat{\mathbf{A}}_s^1[m] - \mathbf{A}_s[m])$ (Alg. 1, Ln. 6), gives an estimate of the fixed-pattern offset due to temporal drift. In the second pass, we subtract this offset from every approximation map and apply spatial processing again, yielding the final approximation coefficients.

4.3.2 Detail Coefficients

The detail coefficients produced by DT-CWT are complex-valued images. We spatially process the real and imaginary parts individually. Note that the spatial processing only encourages spatial coherence and doesn’t itself nullify any detail coefficients. But the temporal noise and quantization in microbolometers introduces a fundamental noise floor. Therefore, we apply soft shrinkage to all detail coefficients

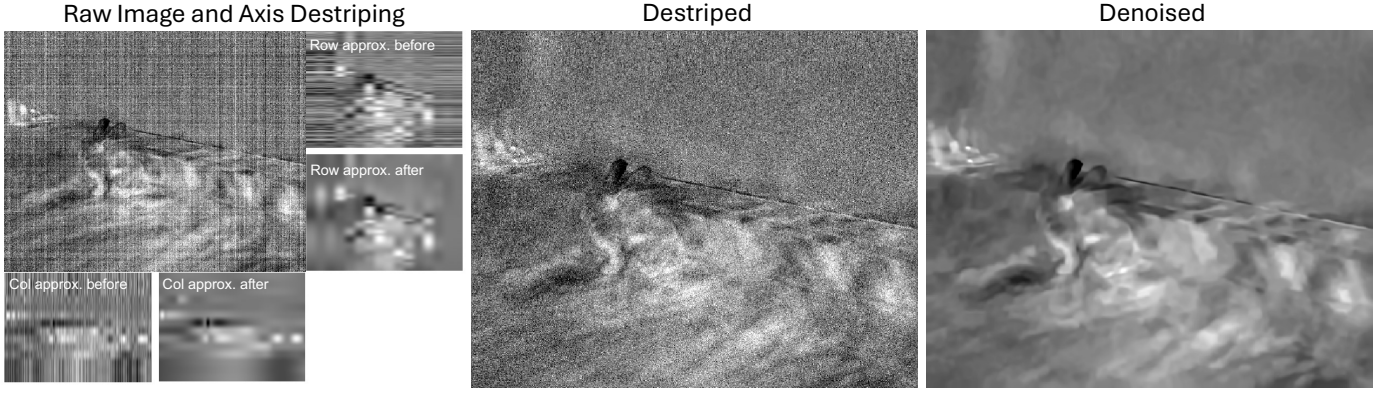


Fig. 7: The real part of the 14th out of 18 detail coefficients at the maximum scale s is shown in the linear gray colormap. To remove vertical stripes, we compute wavelet transform along the column. Each row of approximation coefficients is then smoothed as described in Alg. 2. The approximation coefficients before and after column destriping are shown below the noisy coefficient map. The approximation coefficients along the rows before and after smoothing are shown to the right. The detail coefficients during both stages are left untouched. The resulting image after destriping is shown in the middle. Notice the lack of low frequency spatial banding typical in common thermal destriping algorithms. We then apply an off-the-shelf gaussian denoising algorithm to recover the denoised coefficient map shown on the right.

Algorithm 2 Constrained Axis Destriping

```

1: procedure  $\mathcal{D}(X, \epsilon_{\min})$ 
2:    $C_a, C_d \leftarrow \text{wavedec1D}(X, \text{axis}=0)$ 
3:   for  $i = 1$  to  $M$  do
4:      $\hat{C}_a[i, \cdot] \leftarrow \text{destripe1D}(C_a[i, \cdot], \epsilon_{\min})$ 
5:   end for
6:    $\hat{X} \leftarrow \text{waverec1D}(\hat{C}_a, C_d)$ 
7:   return  $\hat{X}$ 
8: end procedure
9: procedure  $\text{DESTRIPE1D}(\mu, \epsilon_{\min})$ 
10:   $y \leftarrow \text{medianFilter}(\mu)$ 
11:   $\epsilon \leftarrow \max(\text{estimateNoise}(\mu), \epsilon_{\min})$ 
12:   $\hat{\mu} \leftarrow \arg \min_{\mu} \|\Delta\mu\|_2^2 \quad \text{s.t.} \quad y - \epsilon \leq \mu \leq y + \epsilon$ 
13:  return  $\hat{\mu}$ 
14: end procedure

```

Our destriping algorithm $\mathcal{D}(\cdot) : \mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{H \times W}$ (Alg. 2) is applied to both axes as follows:

$$\hat{X} = \mathcal{D}(\mathcal{D}(X)^\top)^\top, \quad (7)$$

where $(\cdot)^\top$ denotes the matrix transpose. The destriper first isolates the stripe noise by computing a 1D wavelet transform along the columns (or rows). Each row (or column) of the approximation coefficients is then smoothed by estimating a trend line using a median filter. The smoothed signal is the one with the minimal Laplacian that stays within an epsilon tube of the trend line. This enables long-range smoothness across the full width (or height) of the image. Figure 7 illustrates the process from noisy input coefficient map to destriped to fully denoised for a detail coefficient from the *bunny draft* scene.

with a universal threshold λ after spatial processing as follows:

$$\hat{\mathbf{D}}_\ell[m] = \text{sign}(\tilde{\mathbf{D}}_\ell[m]) \max(|\tilde{\mathbf{D}}_\ell[m]| - \lambda, 0). \quad (6)$$

We determine the value of λ using the plain wall scene from Fig. 2 as reference. The detail signal for this scene is mostly noise. The median magnitude of its finest detail coefficients was 9.55. Empirically, we found a lower value $\lambda = 6$ preserved more details in all our scenes.

4.3.3 Constrained Axis Destriping and Gaussian Denoising

It is deceptively challenging to find a single spatial processing algorithm for all scenes. In our work, we first apply a destriping algorithm and then use an off-the-shelf Gaussian denoising algorithm on its output. For high SNR scenes, we subtract the axis mean, use a 1D Non-Local Means to smooth it and add this back to construct a destriped image. We repeat for both axes. For low SNR scenes, we develop a custom destriping algorithm described below. In all cases, we use BM3D [11] to denoise the destriped output.

4.4 Mitigating Signal Extension Effects

The DT-CWT requires a boundary extension policy at the start and end of the video. These boundary effects are most visible in the temporal derivatives in low-contrast scenes with mismatched video lengths and at large decomposition depths s , where it introduces temporal ghosting artifacts. To minimize such artifacts, we choose $s = \lfloor \frac{N}{L-1} \rfloor$, where L is the longest filter used in DT-CWT.

4.4.1 Pre-/Post-Transform Steps

For a chosen maximum temporal scale s , we trim the video length to the nearest multiple of 2^s . We also perform global mean detrending as follows

$$\mu[n] = \frac{1}{HW} \sum_i^H \sum_j^W I[n, i, j], \quad \mathbf{I}_{pre} = \mathbf{I} - \mu[n, \cdot, \cdot]. \quad (8)$$

This separates the dominant global temporal mode, such as the entire scene warming up, and reduces the total energy of the signal passed to the DT-CWT.

TABLE 1: Overview of the six thermal signals/phenomena revealed using a noisy low-cost microbolometer.

LWIR Phenomenon	Type of Visualization	Scene Name
Light-to-heat Conversion	Intensity change	Book $\triangleright$ Fig. 8
Heat Conduction	Iso-contour map	Bunny $\triangleright$ Fig. 9
Heat Convection	Temporal derivative	Exhaust $\triangleright$ Fig. 10
Spectral Emission	Temporal derivative	Bulb $\triangleright$ Fig. 11
Polarized intensity	Angular derivative	Bowl $\triangleright$ Fig. 12
Thermal Reflection	Normalized intensity	Yoda $\triangleright$ Fig. 13

We smooth the global mean signal $\mu[n]$ using a cubic B-Spline to obtain $\hat{\mu}[n]$, then add it back after the inverse transform:

$$\hat{\mathbf{I}} = \hat{\mathbf{I}}_{post} + \hat{\mu}[n, \cdot, \cdot]. \quad (9)$$

5 VISUALIZING HEAT FLOWS

Heat transport effects are diverse and depend on the scene, environment and sources of heat. In this section, we consider six types of heat signals/phenomena (summarized in Table 1) that can be observed using noisy low-cost microbolometers. For each phenomenon, we construct a representative scene, demonstrate TSVP, and describe its utility for thermal computational imaging. In all results, the temporal derivative is computed by first applying a gaussian smoothing with $\sigma_t = 1.5$ along the temporal dimension and then taking finite differences.

Light-to-Heat Conversion. Light of any wavelength, upon absorption, is (mostly) converted to heat. This complementary relationship has been used to analytically solve for albedo and shading [5], [6] without data-driven priors. We show one such example in Fig. 8 where a stationary tabletop scene of a book and keyboard heats up as a result of light absorbed from a white LED panel. Subtracting the first frame from subsequent frames eliminates the effect of initial conditions and reveals the differential heating induced by lighting. After TSVP, the text on the book becomes legible and the geometry-induced shading on the keyboard keys is well-preserved.

Heat Conduction. Most Shape-from-X methods recover depth or normals by modeling light transport but cannot directly measure curvature. In contrast, recent work [1], [2] has leveraged heat conduction to directly measure curvature or Shape Laplacians from thermal videos. These slow thermal gradients are imperceptible to the naked eye as our vision is tuned to detecting high contrast moving edges and not slow changes in pixel intensities. A powerful visualization aid is the *wrapped colormap* (iso-contour map) [1]: rather than displaying X directly, one displays $X \bmod r$ with a high-contrast colormap such as *turbo*, where r is user controllable. Figure 9 shows a 3D-printed bunny heated by a nearby incandescent bulb. TSVP enables using smaller values of r , revealing conduction patterns at greater granularity. Downstream shape estimation from this scene is discussed in Sec. 6.2.

Heat Convection. Optical Gas Imaging (OGI) in the thermal infrared is routinely used to image volatile organic compounds and greenhouse gases [42]. Thermal imaging also enables passive detection and tracking of hot or cold air pockets. Figure 10 shows exhaust plumes and radiator

TABLE 2: PSNR and SSIM scores for albedo and shading estimated by JoLHT-Video [5] using the thermal video before and after denoising.

JoLHT-Video	PSNR $\uparrow$		SSIM $\uparrow$	
	Albedo	Shading	Albedo	Shading
w/o denoising	21.95	31.79	0.61	0.87
w/ denoising	24.44(+11%)	32.26(+1.5%)	0.87 (+43%)	0.94 (+8.0%)

airflow from two stationary vehicles, that is only revealed in the temporal derivative computed after TSVP.

Spectral Emission. Like visible light, thermal appearance also varies with wavelength, which has been used in many vision applications [3], [7], [8]. In Fig. 11, we show a scene with two lit incandescent bulbs captured under different LWIR spectral bandpass filters. Despite their intense emission (rated for > 200 W each), the temporal derivative after TSVP with the $10.5 \mu\text{m}$ filter shows the subtle reflections of a person waving in the background as well as the bulbs’ mutual reflections. Presumably due to differences in spectral emissivity, the derivative with the $12.1 \mu\text{m}$ filter only shows the continued heating of the bulb surfaces. The bulbs’ filament is visible in this band but not in the other. The high scene dynamic range makes noise less visible in intensity images, but the temporal derivative is overwhelmed by noise without TSVP.

Polarized Thermal Intensity. A commonly used optical filtering element in LWIR is a wire-grid polarizer. Recent works estimating shape from polarimetric thermal intensity [12], [13] resort to excessively heating the object to improve SNR. In contrast, we show an example scene in Fig. 12 of a glass bowl filled with lukewarm water. The bowl is surrounded by mirror and glass plates that reflect the bowl six times. As the polarizer is rotated, the strengths of the reflections vary. The time derivative of the denoised video shows different reflections clearly while derivatives of the original video are overwhelmed by noise. Scaling the time derivative with the speed of rotation of the polarizer directly gives the angular derivative of the polarized intensity with angle of polarization.

LWIR Surface Reflection. Thermal cameras are sensitive to long-wave infrared light ($8\text{--}14 \mu\text{m}$), which exhibits higher specular reflection than visible light. Extracting these subtle specular reflections is valuable for Non-Line of Sight imaging applications [43]. We show an example in Fig. 13, where the glossy eye region of a doll reflects the thermal radiation from two humans beyond the line of sight. The top row shows in gray scale the clear reduction in both striping and noise using TSVP. The reflections are barely visible in raw data but the denoised frames show the poses of the humans clearly (hands raised and swaying periodically). The intensities at two pixels in the eye region show successful denoising even when the standard deviation of the noisy pixels is over six times the denoised values. The cropped regions have low contrast when using the gray color map, so we use the *twilight* colormap to aid the reader.

6 EXPERIMENTAL VALIDATION

Subtle heat flows are challenging to quantitatively validate in the real world. Our empirical analysis in Sec. 3 shows that

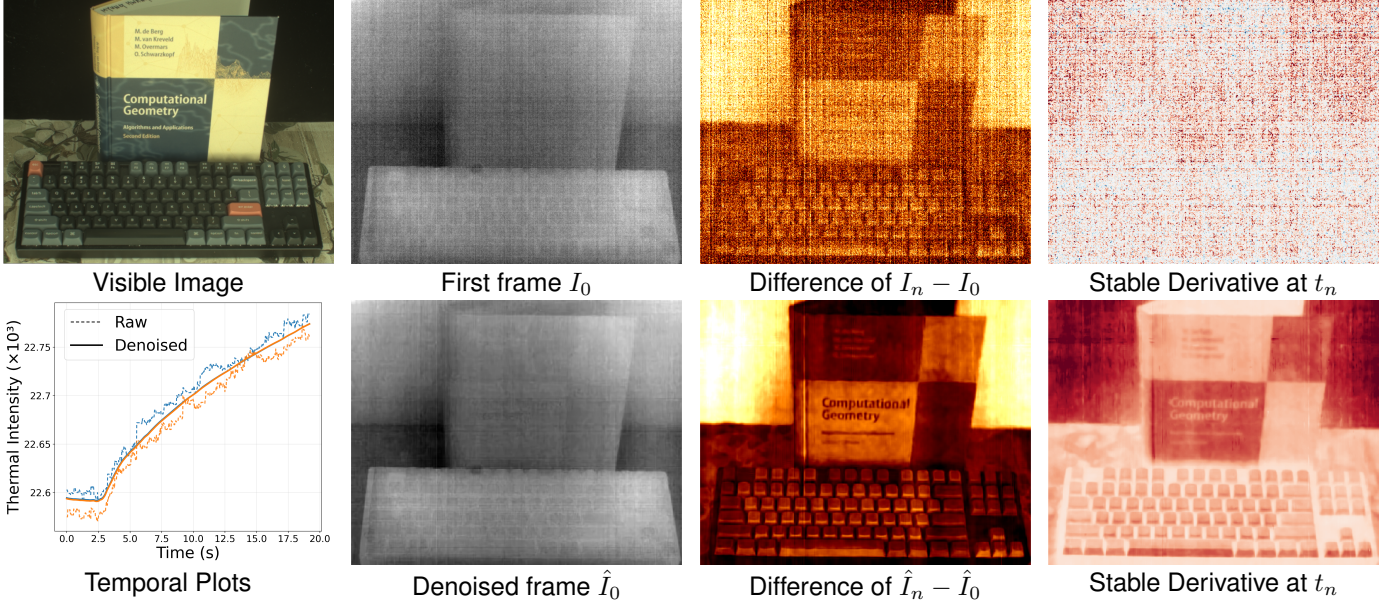


Fig. 8: A tabletop scene of a book and a keyboard front-illuminated by a white LED panel that is turned on 2.5 seconds after the start of the video. Thermal appearance of the first frame is dominated by prevailing ambient conditions. The difference frame reveals the heat accumulated by absorbing light while the temporal derivative reveals the current net rate of heat retention. The writings on the book, the spatially varying albedo dependent heating of the keyboard and the mat beneath are all legibly seen less than a second after the light is turned on (at 3.3 seconds) in both the difference frame and temporal derivative. The temporal plot shows two neighboring pixels on the book surface. The denoised pixel values are on top of each other while the raw pixel intensities are consistently far apart.

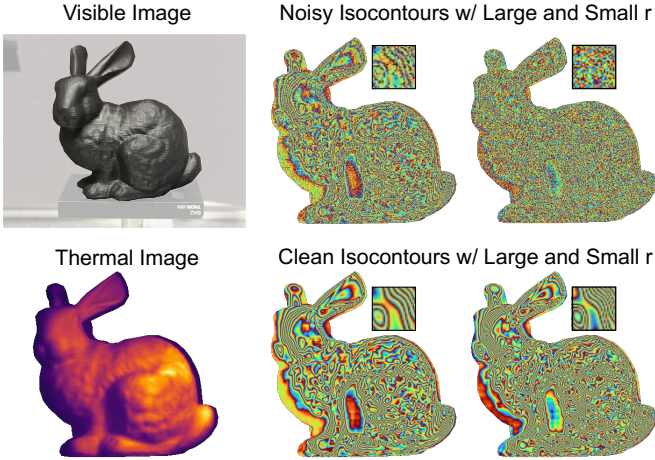


Fig. 9: A 3D printed bunny illuminated by an incandescent light bulb from the right shows heating induced by light absorption. The raw thermal image appears to have good SNR but iso-contour maps computed from it show the destructive impact of noise. The same iso-contour maps computed from our denoised video is robust to even small range values enabling us to visualize heat conduction at greater granularity. Zoom in for best visualization.

common noise models for microbolometers are insufficient for realistic simulation. One option is to benchmark against a cooled thermal camera. Their NETD (45 mK for FLIR A6751 SLS) is not negligible for subtle heat flows (Fig. 1 had median temporal standard deviation of ≈ 200 mK)

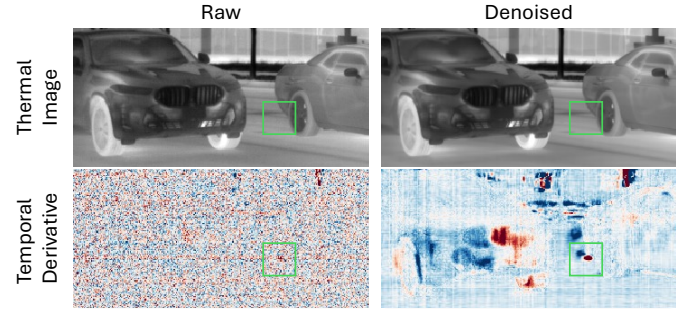


Fig. 10: Two vehicles stationary at a traffic light. Visualizing the temporal derivative after denoising reveals the flow of air across the front radiator of the left vehicle and exhaust out of the right vehicle. The images and derivatives are shown using local contrast enhancement. See supplementary video for air flows.

and would also require denoising. But they are expected to have negligible drift in controlled settings due to their tight temperature stabilization. However, they are expensive ($\approx \$150k$) and difficult to procure. Therefore, we evaluate TSVP through three alternative ways below.

6.1 Consistency under Controlled Experiments

Changing light intensity: We placed a book, a color chart and a light meter in front of a white LED panel and recorded 20 sec videos of the scene heating up at four light levels: L1 (32.6×10^3 lux), L2 (48.4×10^3 lux), L3 (64.7×10^3 lux), and

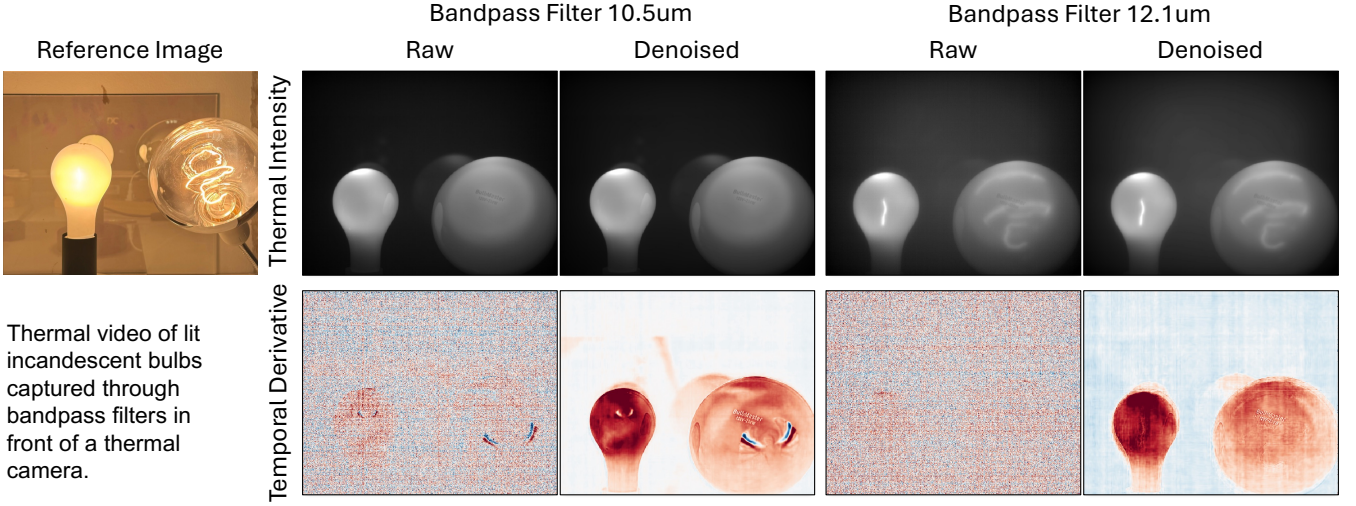


Fig. 11: Multispectral thermal imaging of lit incandescent bulbs using two bandpass filters — CWL $10.5\mu\text{m}$ (FWHM $1.5\mu\text{m}$) and CWL $12.1\mu\text{m}$ (FWHM $0.5\mu\text{m}$) — placed in front of a thermal camera. Both raw and denoised thermal intensity images are shown along with their temporal derivatives. The $10.5\mu\text{m}$ filter captures broad thermal emission from the bulbs; its temporal derivative reveals subtle reflections of a person waving in the background, despite the high foreground temperatures. The narrower $12.1\mu\text{m}$ filter resolves the lamp filament in the thermal intensity image, while its temporal derivative highlights the bulbs’ gradual heating but suppresses the background motion entirely, suggesting stronger spectral selectivity at this wavelength.

L4 (76.4×10^3 lux). From the theory in [5][Eq. 9, 13], the rise in temperature for a short duration after illumination obeys:

$$I_n^l - I_0^l = \frac{\eta^l(1 - \rho)k_1\alpha}{P}(1 - e^{-\frac{t_n - t_0}{H/P}}), \quad (10)$$

where l denotes the light level and other variables are as in [5]. Since the scene is static, only the shading term η^l scales between experiments. While we cannot directly measure per-pixel irradiance, the *ratio* of temperature rises at two light levels will equal the ratio of light levels:

$$\frac{I_n^{L_4} - I_0^{L_4}}{I_n^{L_i} - I_0^{L_i}} = \frac{L_4}{L_i}, \quad \forall i \in \{1, 2, 3\}. \quad (11)$$

Figure 14 shows that the ratio maps computed from the denoised videos are nearly uniform even though the ratio operator amplifies noise. The top row shows the intensity change images for frame $n = 200$ with and without denoising. The correlation plot in the bottom right shows that the mean estimated ratios from the denoised videos are closer to ground truth with a significantly smaller standard deviation than those computed from noisy input videos.

Changing fan speed: We heated a thick black cardboard with a 250 W heat lamp for several minutes, then recorded 20 sec videos of its cooling by a PC fan running at different speeds (Fig. 15). The cardboard is allowed to heat back up before recording with a new fan speed. The computed temporal derivatives visualize the spatially varying heating/cooling due to wind motion. Because convection is fundamentally chaotic, the derivatives are not directly comparable pixel-wise across runs. But the patterns show that the direction of wind motion is consistent with the location of the fan across different fan speeds. We also notice that increasing the fan speed results in more laminar flows with high frequency patterns.

6.2 Quantitative Impact on Downstream Tasks

We consider three downstream tasks that use a thermal video as input and demonstrate improvement in their performance after applying TSVP.

Intrinsic Image Decomposition (IID). JoLHT-Video [5] uses a thermal video of the scene after illumination and a single visible image to estimate its albedo and shading. As seen in the pixel plot in Fig. 17, the exponential curve fit (blue curve) has a large discrepancy with the raw measurements (blue dots). However, the denoised video (green dots) is largely monotonic, more closely resembles a plausible thermal transient and results in a better exponential curve fit (green curve). The spatial coherence enforced by TSVP directly results in quantitative improvements in the accuracy of albedo and shading (shown in Tab. 2) using the same pixel-wise algorithm.

Shape from heat conduction. Recall from Fig. 9 that the denoised video affords visualizing conduction (iso-contour maps) at greater granularity. In Fig. 16, we see that the absorbed light intensity estimated from the denoised video shows greater geometric detail. The shape computed from denoised videos shows a better curvature of the bunny’s belly. The resulting shape reconstruction using [2] has a reduced Median Absolute Error (MAE) of 4.98 mm from 5.62 mm.

Thermal reflection and emission separation. We apply the dual-band separation algorithm of [7] to their *coffee pot* scene. Note that their algorithm includes internal smoothing and noise estimation. As seen in Fig. 18, pre-processing with TSVP before applying their algorithm yields a more visually consistent result.

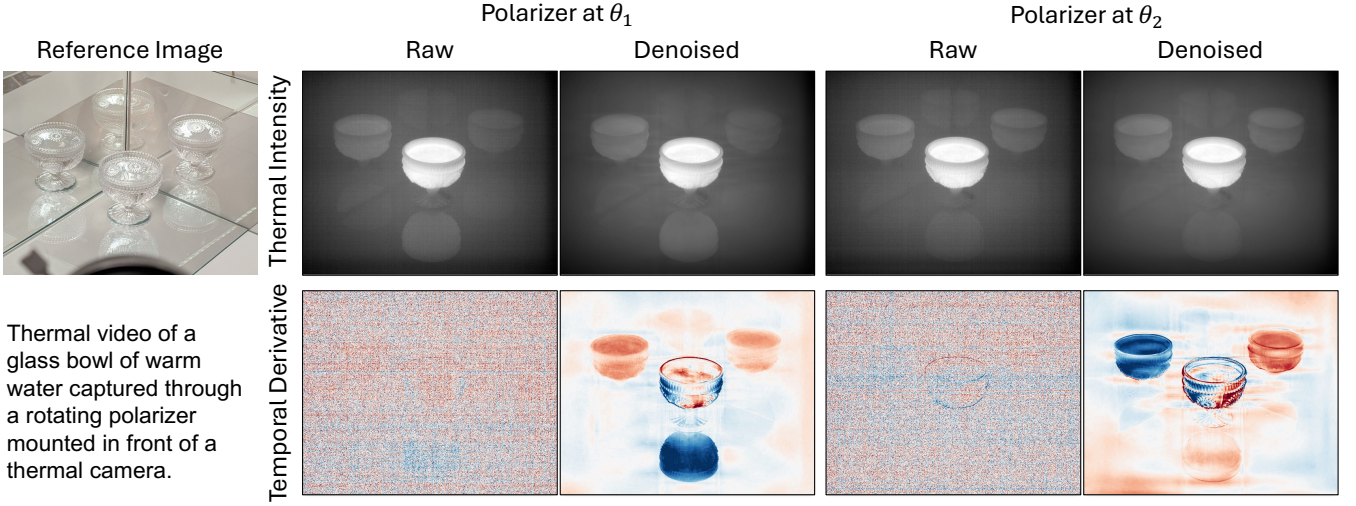


Fig. 12: Polarimetric thermal imaging of a glass bowl filled with lukewarm water. A total of 7 visual copies of the bowl including real and reflected are seen. As the polarizer is rotated, the temporal derivative highlights different reflections as they come in and go out of phase with the polarizer. These could be used to extend vision algorithms that leverage polarization cues to the thermal spectrum.

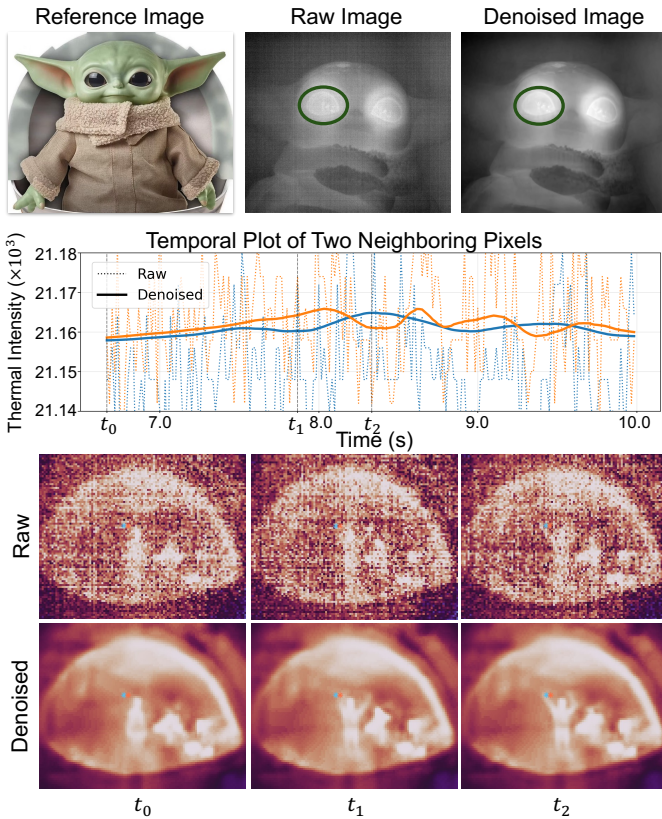


Fig. 13: The glossy eyes of a doll is specular in LWIR and reflects thermal radiation from two humans beyond the line of sight. The human seen on the left has their hands raised and moving side-to-side. The temporal plot of two pixels alternatively peaks as the hands switches positions between the pixels. The standard deviation of the noisy pixel values is over $6\times$ that of the denoised values. The cropped regions are shown in *twilight* colormap for better contrast between the human and the background.

6.3 Qualitative Comparison with Baselines

We compare TSVP against five baselines on the challenging *bunny draft* scene (Fig. 19), where a gentle draft cools the scene by < 0.2 K (median temporal standard deviation).

BM3D [11] is a state-of-the-art model-based denoiser for additive Gaussian noise. While the denoised intensity appears smooth, the temporal derivative is corrupted by stripe noise.

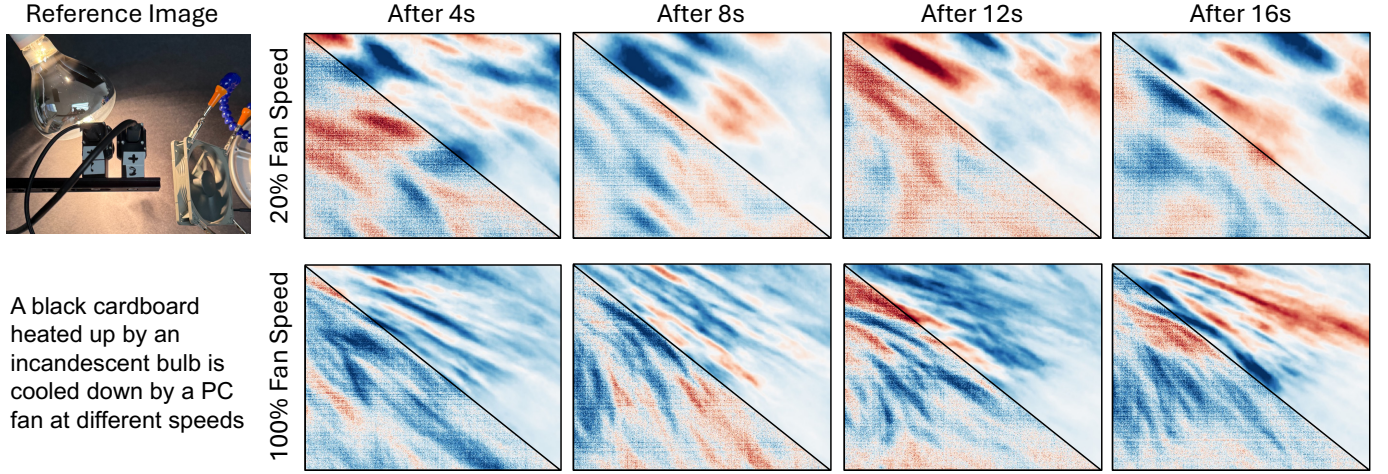
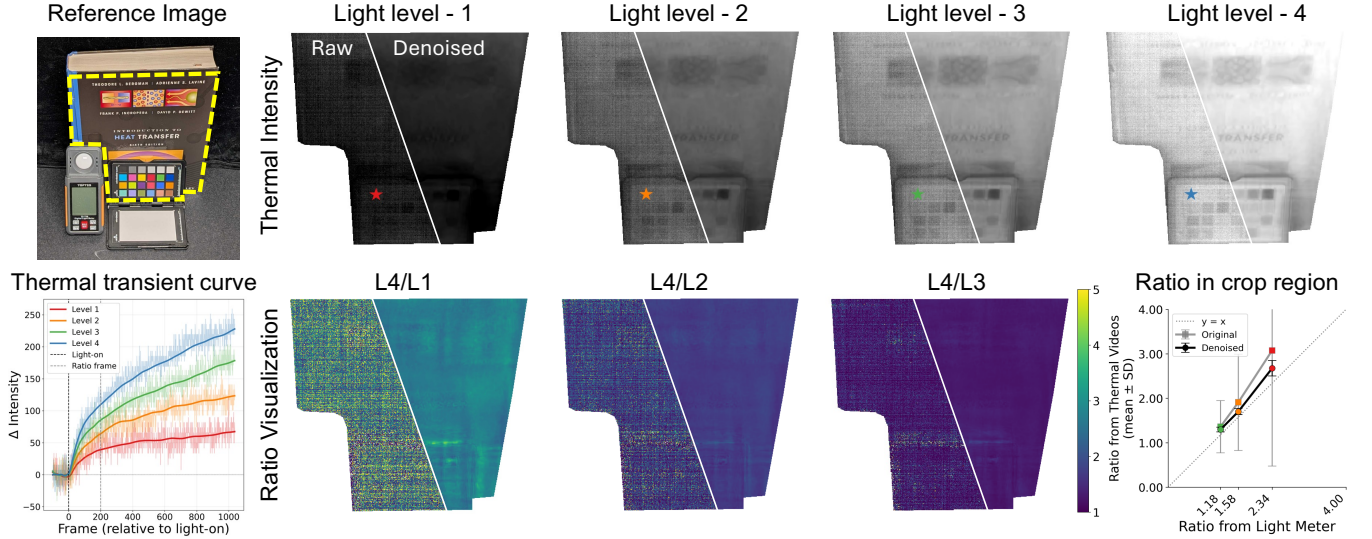
On-board FLIR Boson+ [44]: we used a second identical thermal camera with their spatio-temporal signal processing algorithms enabled. While the stripes are removed in the intensity images, the temporal derivative still lacks spatial structure.

DeepIR [21] is an optimization based algorithm that relies on camera jitter to process a stack of thermal images and output a single clean thermal image. Since this scene is static, the spatial structures get conflated with the jointly estimated gain and offset maps by the algorithm. As it outputs a single image, temporal derivative is not applicable for this method.

DEAL [31] is a learning-based single-image denoising algorithm for thermal cameras. Its output contains consistent spatial artifacts, and the temporal derivative shows no plausible scene structure.

MDIVDnet [26] is a learning-based thermal video restoration algorithm. We could not fit the entire 1200-frame video in memory. When we provided the 21 frames centered around the target, there were no legible spatial structures in the temporal derivative. So we instead uniformly subsampled frames from the 1200-frame video to provide full context. The temporal derivative looks plausible but corresponds to the coarse temporal derivative and lacks the detailed wind-motion patterns.

Frame-wise Spatial Processing (our destriper + BM3D without temporal transform): the temporal derivative has fewer stripe artifacts than BM3D alone but still fails to reveal the highly detailed wind-motion patterns seen in our



TSVP result. This ablation directly confirms the temporal transform is the critical ingredient.

7 DISCUSSION

Role of thermal time constant. When a bolometric pixel with time constant < 10 ms is sampled at 60 Hz (16.67 ms), a step change in incident flux gets smeared across 2 – 3 frames implying that our target signal is highly likely to be continuous with discontinuities in the first derivative.

Using low FPS videos. Temporally sub-sampling a microbolometer video will introduce aliasing. Applying TSVP

to such sub-sampled videos is erroneous even if it produces plausible videos. Given that the primary motivation for working with low FPS videos is computational efficiency, we observe that the sparsity in the temporal coefficient domain could enable efficient vision algorithms that operate at the native frame rate of the microbolometer itself.

Learning-based spatial processors. We tested using DnCNN [45] in the place of BM3D but with little success. However, developing novel learning-based denoising algorithms for the coefficient domain is a promising direction for future work.

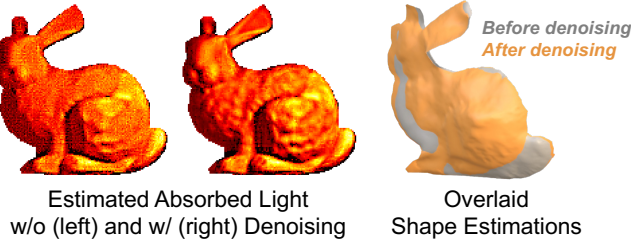


Fig. 16: Qualitative results of shape reconstruction by Shape From Laplacian and Shading [2] method using raw video and the denoised video. The estimated heat source intensity shows greater geometric detail on the body of the bunny, its toes and ears. The resulting shape has higher curvature that is visibly closer to the ground truth.

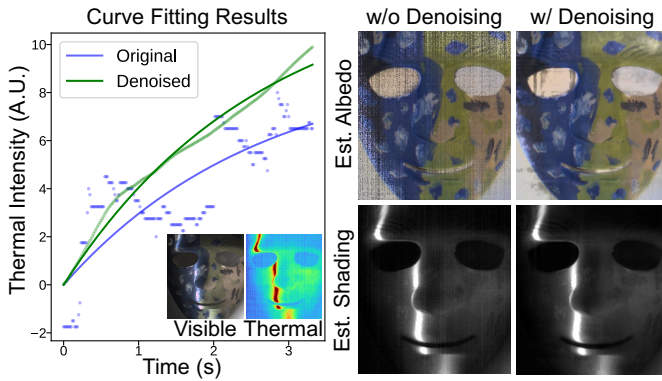


Fig. 17: Albedo-shading separation using JoLHT-Video [5] before and after denoising with our algorithm. The blue dots from the raw video hardly shows the exponential-like behavior expected from the heat transport equation. The green dots from the denoised video is more representative of the physics of heat transport even though our denoising does not explicitly model any physical priors. The denoised video results in a spatially coherent albedo-shading separation without resorting to strong structural priors as in [6].

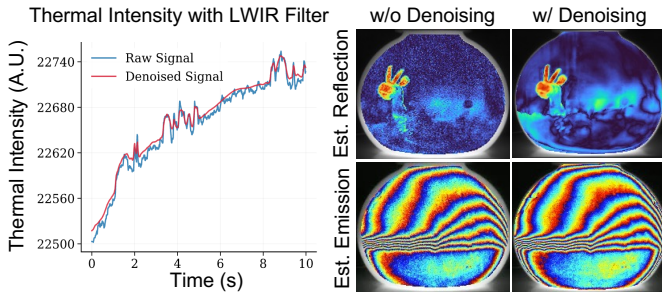
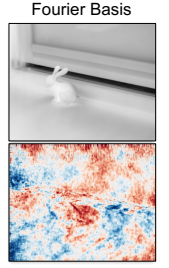


Fig. 18: The impact of denoising before reflection-emission separation using Dual-band Thermal Videography [7]. The temporal plot on the left shows the signal with a narrow bandpass filter CWL $9.5 \mu\text{m}$ (FWHM $0.5 \mu\text{m}$) in front of the thermal camera before and after denoising. The estimated reflection component from the denoised video is cleaner and shows greater detail than the original result.

Why not use Fourier Transform along time?

Naively speaking, it is tempting to use the computationally simpler Fourier basis for the temporal transform. However, since Fourier transform has no time localization, a sharp temporal edge, such as due to a light turning on or the onset of a wind draft, gets smeared across all coefficients in the frequency domain. As seen here, spatially smoothing these coefficients destroys this timing information even though the recovered intensity images look spatially smooth.



What about other temporal transforms? We chose DT-CWT based on our desired transform properties but we do not claim it to be optimal. Designing better temporal transforms, perhaps in conjunction with an optimal spatial processor are interesting directions for future work.

8 CONCLUSION

Imaging heat flows can provide us with a richer understanding of the world around us. We demonstrated that properly modeling the bolometric pixel dynamics allows us to extract subtle heat transport phenomena from noisy, low-contrast, and low-cost microbolometric videos. Our wavelet-based approach is simple, highly parallelizable and the visualizations are compelling. Our denoised videos improve the performances of several downstream tasks and can enable future thermal computational imaging approaches that rely on a variety of optical filtering. Like the motion magnification work that revealed subtle motions [46] from ordinary RGB videos, we hope our work inspires new research on thermal imaging to understand thermodynamics in complex everyday scenes.

ACKNOWLEDGMENTS

This work was partly supported by NSF grants IIS-2107236, and NSF-NIFA AI Institute for Resilient Agriculture. Kao Corporation supported Akihiko Oharazawa's effort. We would also like to thank Aswin C. Sankaranarayanan for helpful discussions.

REFERENCES

- [1] S. Narayanan, M. Ramanagopal, M. Sheinin, A. C. Sankaranarayanan, and S. G. Narasimhan, "Shape from heat conduction," in *European Conference on Computer Vision*. Springer, 2024, pp. 426–444.
- [2] A. Oharazawa, S. Narayanan, M. Ramanagopal, and S. G. Narasimhan, "Resolving shape ambiguities using heat conduction and shading," in *Proceedings of the IEEE/CVF International Conference on Computational Photography (ICCP)*, July 2025, pp. 182–192.
- [3] Y. Nagase, T. Kushida, K. Tanaka, T. Funatomi, and Y. Mukaigawa, "Shape from thermal radiation: Passive ranging using multi-spectral lwir measurements," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 12 661–12 671.
- [4] A. Dashpute, V. Saragadam, E. Alexander, F. Willomitzer, A. Katsaggelos, A. Veeraraghavan, and O. Cossairt, "Thermal spread functions (tsf): Physics-guided material classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 1641–1650.

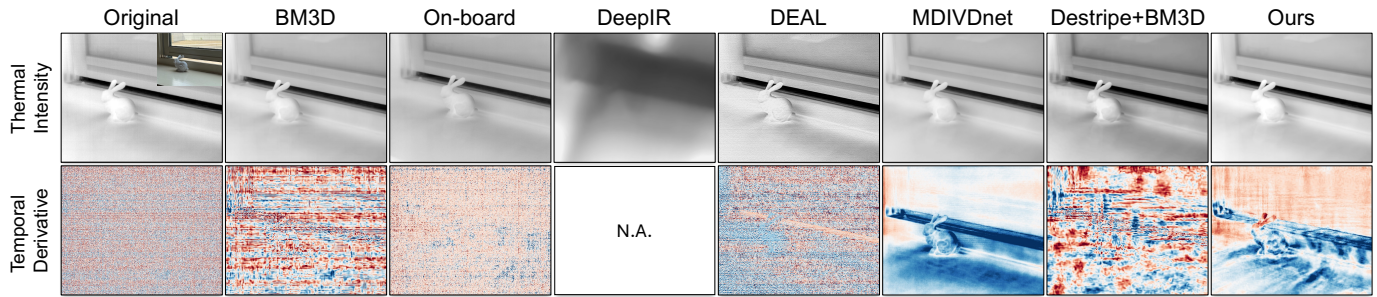


Fig. 19: Qualitative comparison of the denoised video and its temporal derivative with several baselines for the *bunny draft* scene (a bunny is near an office window with gentle breeze flowing in). BM3D [11] is designed for gaussian noise and is unable to remove stripe noise. The on-board signal processing algorithms suppress stripe noise but its temporal derivative is overwhelmed by noise. DeepIR [21] requires camera jitter which is absent in this scene. It conflates scene content with the jointly estimated gain and offset maps. As it predicts a single image given a video, temporal derivative is not applicable. DEAL [31] is a learning-based single-image restoration algorithm designed for thermal cameras, but suffers from strong artifacts. MDIVDnet [26] is a learning-based thermal video restoration algorithm. With 21 frames subsampled from the 1200-frame video, it recovers the coarse temporal derivative which lacks the high frequency details of the wind movement that is clearly seen in our denoised visualizations. Applying our proposed destripe+BM3D spatial processing to each frame in the time domain also fails, further emphasizing the need to operate in the temporal coefficient domain.

- [5] M. Ramanagopal, S. Narayanan, A. C. Sankaranarayanan, and S. 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)*, June 2024, pp. 11 924–11 933.
- [6] Z. Yuan, M. Ramanagopal, A. C. Sankaranarayanan, and S. G. Narasimhan, "Vt-intrinsic: Physics-based decomposition of reflectance and shading using a single visible-thermal image pair," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.
- [7] S. Narayanan, M. Ramanagopal, and S. Narasimhan, "Dual band thermal videography: Separating time-varying reflection and emission near ambient conditions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.
- [8] F. Bao, X. Wang, S. H. Sureshbabu, G. Sreekumar, L. Yang, V. Aggarwal, V. N. Boddeti, and Z. Jacob, "Heat-assisted detection and ranging," *Nature*, vol. 619, no. 7971, pp. 743–748, 2023. [Online]. Available: <https://doi.org/10.1038/s41586-023-06174-6>
- [9] N. Kingsbury, "The dual-tree complex wavelet transform: a new efficient tool for image restoration and enhancement," in *9th European signal processing conference (EUSIPCO 1998)*. IEEE, 1998, pp. 1–4.
- [10] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 11, no. 7, pp. 674–693, 2002.
- [11] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.
- [12] T. Kushida and K. Tanaka, "Thermal polarimetric multi-view stereo," *arXiv preprint arXiv: 2510.20972*, 2025.
- [13] K. Kitazawa and T. Takatani, "Shape from polarization of thermal emission and reflection," *International Conference on Computational Photography*, 2025.
- [14] R. Liu and C. Vondrick, "Humans as light bulbs: 3d human reconstruction from thermal reflection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 12 531–12 542.
- [15] M. Sheinin, A. Sankaranarayanan, and S. G. Narasimhan, "Projecting trackable thermal patterns for dynamic computer vision," in *Proc. IEEE/CVF CVPR*, 2024.
- [16] S. N. Torres, J. E. Pezoa, and M. M. Hayat, "Scene-based nonuniformity correction for focal plane arrays by the method of the inverse covariance form," *Applied optics*, vol. 42, no. 29, pp. 5872–5881, 2003.
- [17] R. C. Hardie, M. M. Hayat, E. Armstrong, and B. Yasuda, "Scene-based nonuniformity correction with video sequences and registration," *Applied optics*, vol. 39, no. 8, pp. 1241–1250, 2000.
- [18] R. C. Hardie and D. R. Droege, "A map estimator for simultaneous superresolution and detector nonuniformity correction," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 089354, 2007.
- [19] Y. Tundero and J. Gilles, "Admire: a locally adaptive single-image, non-uniformity correction and denoising algorithm: application to uncooled ir camera," in *Infrared Technology and Applications XXXVIII*, vol. 8353. SPIE, 2012, pp. 580–595.
- [20] Z. He, Y. Cao, Y. Dong, J. Yang, Y. Cao, and C.-L. Tisse, "Single-image-based nonuniformity correction of uncooled long-wave infrared detectors: A deep-learning approach," *Applied optics*, vol. 57, no. 18, pp. D155–D164, 2018.
- [21] V. Saragadam, A. Dave, A. Veeraraghavan, and R. G. Baraniuk, "Thermal image processing via physics-inspired deep networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4057–4065.
- [22] A. Barral, P. Arias, and A. Davy, "Fixed pattern noise removal for multi-view single-sensor infrared camera," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1669–1678.
- [23] A. Wolf, J. E. Pezoa, and M. Figueroa, "Modeling and compensating temperature-dependent non-uniformity noise in microbolometer cameras," *Sensors*, vol. 16, no. 7, p. 1121, 2016.
- [24] M. Maggioni, E. Sánchez-Monge, and A. Foi, "Joint removal of random and fixed-pattern noise through spatiotemporal video filtering," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4282–4296, 2014. [Online]. Available: <https://doi.org/10.1109/TIP.2014.2345261>
- [25] M. Tassano, J. Delon, and T. Veit, "Fastdvdnet: Towards real-time deep video denoising without flow estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [26] L. Cai, X. Dong, K. Zhou, and X. Cao, "Exploring video denoising in thermal infrared imaging: Physics-inspired noise generator, dataset, and model," *IEEE Transactions on Image Processing*, vol. 33, pp. 3839–3854, 2024.
- [27] X. Kuang, X. Sui, Y. Liu, Q. Chen, and G. Gu, "Single infrared image optical noise removal using a deep convolutional neural network," *IEEE photonics journal*, vol. 10, no. 2, pp. 1–15, 2017.
- [28] P. Xiao, Y. Guo, and P. Zhuang, "Removing stripe noise from infrared cloud images via deep convolutional networks," *IEEE Photonics Journal*, vol. 10, no. 4, pp. 1–14, 2018.
- [29] J. Guan, R. Lai, A. Xiong, Z. Liu, and L. Gu, "Fixed pattern noise reduction for infrared images based on cascade residual attention cnn," *Neurocomputing*, vol. 377, pp. 301–313, 2020.
- [30] J. Guan, R. Lai, and A. Xiong, "Wavelet deep neural network for stripe noise removal," *IEEE Access*, vol. 7, pp. 44 544–44 554, 2019.
- [31] Z. Liu, Z. Wang, J. Liu, F. Meng, L. Ma, and R. Liu, "Deal: Data-efficient adversarial learning for high-quality infrared imaging," in

Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 28 198–28 207.

- [32] T. H. Rhee, D.-g. Lee, and A. Kim, “Tir-diffusion: Diffusion-based thermal infrared image denoising via latent and wavelet domain optimization,” *arXiv preprint arXiv:2508.03727*, 2025.
- [33] Y. Liu, Z. Wang, L. Si, L. Zhang, C. Tan, and J. Xu, “A non-reference image denoising method for infrared thermal image based on enhanced dual-tree complex wavelet optimized by fruit fly algorithm and bilateral filter,” *Applied Sciences*, vol. 7, no. 11, p. 1190, 2017.
- [34] J.-Z. Wu, J.-J. Lin, Z.-J. Gao, H.-H. Lin, Q.-Q. Li, and C.-C. Zhang, “A hybrid fourier-wavelet denoising method for infrared image of porcelain sleeve cable terminal using gsm model for wavelet coefficients,” in *2018 International Conference on Power System Technology (POWERCON)*. IEEE, 2018, pp. 1919–1923.
- [35] M. Li, Y. Wang, and H. Sun, “Single-frame infrared image non-uniformity correction based on wavelet domain noise separation,” *Sensors*, vol. 23, no. 20, p. 8424, 2023.
- [36] H. Xu, L. Dong, M. Wang, Y. Ji, and F. Tang, “A real maritime infrared image denoising network based on joint spatial and wavelet domains,” *Journal of Marine Science and Engineering*, vol. 14, no. 7, p. 644, 2026.
- [37] H. Li, Y. Zhang, L. Yang, and H. Zhang, “Infrared image denoising algorithm based on wavelet transform and self-attention mechanism,” *Sensors*, vol. 26, no. 2, p. 523, 2026.
- [38] L. Liu, L. Xu, and H. Fang, “Simultaneous intensity bias estimation and stripe noise removal in infrared images using the global and local sparsity constraints,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 1777–1789, 2019.
- [39] M. Bouali and S. Ladjal, “Toward optimal destriping of modis data using a unidirectional variational model,” *IEEE Transactions on geoscience and remote sensing*, vol. 49, no. 8, pp. 2924–2935, 2011.
- [40] M. Alhussein and S. I. Haider, “Simulation and analysis of uncooled microbolometer for serial readout architecture,” *Journal of Sensors*, vol. 2016, no. 1, p. 9751056, 2016.
- [41] R. Wareham, N. Kingsbury, and C. Shaffrey, “dtcwt: A python implementation of the dual-tree complex wavelet transform,” <https://dtcwt.readthedocs.io/en/0.12.0/>, 2013, accessed: 2026-06-13.
- [42] M. Vollmer and K.-P. Möllmann, *Fundamentals of Infrared Thermal Imaging*. John Wiley & Sons, Ltd, 2017, ch. 1, pp. 1–106. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527693306.ch1>
- [43] T. Maeda, Y. Wang, R. Raskar, and A. Kadambi, “Thermal non-line-of-sight imaging,” in *2019 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2019, pp. 1–11.
- [44] Teledyne FLIR, “Boson+ longwave infrared thermal camera datasheet,” <https://www.flir.com/products/boson-plus/>, 2024, teledyne FLIR LLC. Accessed: 2026-06-13.
- [45] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising,” *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
- [46] M. Rubinstein, N. Wadhwa, F. Durand, and W. T. Freeman, “Revealing invisible changes in the world,” *Science*, vol. 339, no. 6119, 2013.



Mani Ramanagopal is a Project Scientist at the Robotics Institute at Carnegie Mellon University. He obtained his Ph.D. in Robotics from the University of Michigan in 2022 and a B.Tech & M.Tech in Electrical Engineering from Indian Institute of Technology Madras in 2016. He has received a Best Paper Award from IEEE RA-L journal in 2023 and is the recipient of the Richard F. and Eleanor A. Towner Prize for Distinguished Academic Achievement for Outstanding Graduate Student. His research interests are

in robotics, vision, and imaging, with a focus on physics-based approaches.



Akihiko Oharazawa is a visiting researcher at the Robotics Institute, Carnegie Mellon University, where he works with Prof. Srinivasa Narasimhan. He is also a Research Scientist at the Skincare Laboratory of Kao Corporation. He received his M.S. in Comprehensive Human Science from the University of Tsukuba. His current research interests lie in physics-based vision, computer vision, and computational imaging.



Sriram Narayanan is a Ph.D. student in the Robotics Institute at Carnegie Mellon University, advised by Prof. Srinivasa Narasimhan. His research lies at the intersection of computer vision, computational imaging, and robotics, with a focus on physics-based methods for heat and light transport. Prior to CMU, he was a Research Scholar at NEC Laboratories America. He is a recipient of the Uber Presidential Fellowship.



Zeqing Yuan is a graduate student in the Robotics Institute at Carnegie Mellon University, advised by Prof. Srinivasa Narasimhan and Prof. Aswin Sankaranarayanan. Prior to CMU, he obtained his bachelor's degree in Computer Science with the highest honor from Zhejiang University. His research lies in computer vision, computational imaging and robotics, with a recent focus on vision with light and heat.



Srinivasa G. Narasimhan is the Interim Director and U. A and Helen Whitaker Professor of the Robotics Institute at Carnegie Mellon University. He obtained his Ph.D. from Columbia University. His group focuses on novel techniques for imaging and illumination to enable applications in vision, graphics, robotics, agriculture, intelligent transportation, and medical imaging. His works have received more than a dozen Best Paper or Best Demo or Honorable mention awards at major conferences. In addition, he has received

the Ford URP Award (2013), Okawa Research Grant (2009) and the NSF CAREER Award (2007). He is the co-inventor of programmable headlights, Aqualux 3D display, Assorted pixels, Motion-aware cameras, Episcan360, Episcan3D, EpiToF3D, and programmable triangulation light curtains. He served on the editorial board of the International Journal of Computer Vision and serves as Senior or Lead Area Chair of top computer vision conferences.