Thermal Image Processing via Physics-Inspired Deep Networks

Vishwanath Saragadam, Akshat Dave, Ashok Veeraraghavan, and Richard G. Baraniuk

Rice University, Houston TX

{vishwanath.saragadam, akshat.dave, vashok, richb}@rice.edu

Abstract

*We introduce **DeepIR**, a new thermal image processing framework that combines physically accurate sensor modeling with deep network-based image representation. Our key enabling observations are that the images captured by thermal sensors can be factored into slowly changing, scene-independent sensor non-uniformities (that can be accurately modeled using physics) and a scene-specific radiance flux (that is well-represented using a deep network-based regularizer). DeepIR requires neither training data nor periodic ground-truth calibration with a known black body target–making it well suited for practical computer vision tasks. We demonstrate the power of going DeepIR by developing new denoising and super-resolution algorithms that exploit multiple images of the scene captured with camera jitter. Simulated and real data experiments demonstrate that DeepIR can perform high-quality non-uniformity correction with as few as three images, achieving a 10dB PSNR improvement over competing approaches.*

1. Introduction

Long wave infrared (LWIR) thermal cameras capture a scene's intensity in the wavelengths spanning 8–14μm. Thermal cameras in LWIR wavelengths find important applications in various scenarios including autonomous driving [1], robust computer vision [2–4], and large scale temperature monitoring [5]. This democratization of thermal imaging is enabled by advances in low-cost uncooled microbolometer sensors. Despite the wide range of applications, uncooled microbolometer sensors face some unique challenges. First, due to sensor-specific noise properties such as non-uniform per-pixel gain and high readout noise the signal to noise ratio is often low. Second, the internal heating of the camera creates "self-imaging," artifacts called the *narcissus effect* [6]. It is hence imperative to augment the low-cost sensors with effective hardware and software solutions to produce high quality images.

There have been several approaches to enhance thermal images by combining multiple measurements [7, 8], using
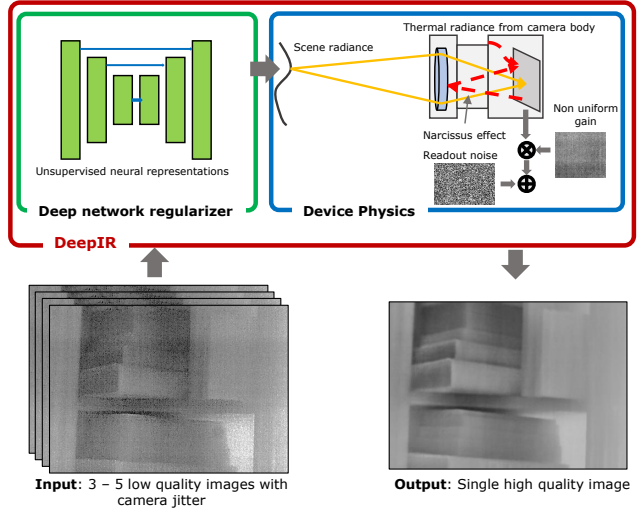

Figure 1: **DeepIR thermal image processing.** DeepIR is a novel thermal camera processing pipeline that combines physically accurate sensor modeling with deep networks to solve inverse problems in the thermal domain. We rely on capturing multiple, jittered images of the scene and then simultaneously estimates the scene's radiant flux by regularizing with a deep network-based regularization.

data driven models [9], and multi modal fusion [10]. Most approaches rely on strong assumptions about the spatial distribution of noise, such as the noise being unbiased or that the noise affects all pixels equally in a column. In real images, such models do not completely capture the statistics of the noise, inevitably leading to poor recovery, or requiring dozens of images to produce high quality images.

A key observation about uncooled thermal sensors is that the image of a scene can be factored into a scene independent component and a scene dependent component (see Fig. 2). The scene independent component includes the gain and offset which arise due to slowly changing thermal conditions within the camera. The scene dependent component includes the scene's radiant flux. We exploit this observation by capturing multiple images of scene with camera motion which only affects the scene's radiant flux measure-
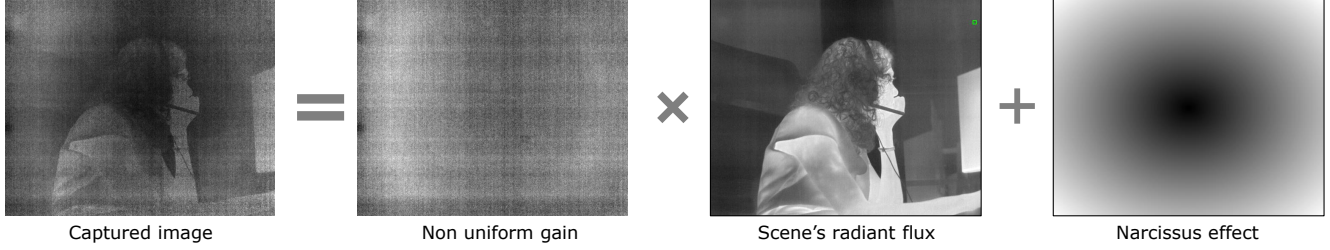
Figure 2: **Non-uniform noise in microbolometer thermal cameras.** This figure visualizes a simulation of image formation with a microbolometer sensor. Due to thermal changes within the camera, the final measurement suffers from spatially varying gain and offset.

ment and not the camera non-uniformities. We then estimate the camera non-uniformities and the scene's radiant flux with a joint optimization approach. To solve the inverse problem, we rely on the regularizing capabilities of convolutional neural network [11] which provides a concise representation for the scene's radiant flux.

The culmination of our efforts is a new image processing pipeline that we call *DeepIR* (pronounced "deeper"), for Deep InfraRed image processing. DeepIR can be used for recovering high quality images from a very small set of images captured with camera motion. We demonstrate the advantages of DeepIR through several simulated and real experiments including non-uniformity correction, super resolution, and narcissus effect suppression. An overview of the DeepIR pipeline is shown in Fig. 1.

The rest of the paper is organized as follows. We review the relevant prior work in section 2 and the physics of uncooled microbolometer sensors in 3. This motivates our multi-frame measurement strategy explained in section 4. We then dive DeepIR into image enhancement with deep network-based representation in section 5, and compare against prior art in 6. We conclude in section 7 with some notes on future directions. To enable further research in thermal image processing, we have made our source code and datasets publicly available[1].

2. Prior Work

Thermal cameras are based either on photonic sensors or microbolometers. Photonic sensors rely on semiconductors to absorb light photons, whereas microbolometers utilize a temperature-dependent resistance to convert thermal radiation to digital output. Due to low manufacturing costs and no external cooling, microbolometer cameras are cheap and compact – making them amenable for several vision-based tasks. We hence focus on microbolometer-based cameras throughout the paper.

Most low-cost microbolometer cameras do not employ thermal stabilization of the focal plane array (FPA), making the measurements highly sensitive to temperature

changes. This results in a slowly drifting non-uniformity that degrades the quality of the image (see Fig. 2). It is hence important to correct for the sensor-specific non-uniformities to obtain accurate measurements. Methods for non-uniformity correction (NUC) for microbolometer sensors can be broadly categorized as hardware-based or software-based.

Hardware approaches. NUC can be performed reliably with an image of flat blackbody at a known temperature. The most popular solution in this approach is the so-called shutter-based flat field (FFC) which relies on periodically capturing images with a closed shutter. Such approaches are not ideal as the mechanical components induce vibrations, and significantly increase power consumption. Solutions which involve a semi-transparent shutter have been proposed that remove the necessity to close the camera [12] but require extremely careful calibration of output reference for each operating temperature.

Software approaches. These exploit the unique properties of microbolometer to correct for non-uniformities, either using a single image [9, 13] or multiple images [7, 8, 14]. Of particular interest in this regard is the work by Hardie et al. [7, 8] which models the image formation as a product of fixed camera-specific gain, and a moving scene-specific radiance. Parameters are then estimated by solving a simple least squares problem. While the approach is promising, the estimated image is sensitive to accuracy of registration and the initial estimate.

DeepIR is inspired by the works of Hardie et al. [7, 8] that combines multiple images of a scene captured with camera motion. Our core contribution is an end-to-end pipeline that jointly estimates the camera non-uniformities, and the scene's radiant flux. We achieve this by regularizing the inverse problem with a concise deep prior-based image representation.

3. Physics of Microbolometer Sensors

Our goal is to recover a high quality image for a few, low quality thermal images corrupted by non-uniform noise. We first present a simple image formation model which moti-

---

[1] https://github.com/vishwa91/DeepIR

vates the DeepIR image processing pipeline.

Sensor modeling. Consider a single pixel in the 2D sensor. Let Φ_{scene} be the radiant flux incident on the pixel and Φ_{fpa} be flux emitted by the pixel. Let C be the thermal capacitance of the microbolometer pixel, and G its thermal conductance. The resulting change in temperature ΔT is related to the above quantities by the energy conservation equation [6]

$$\alpha(\Phi_{\text{scene}} - \Phi_{\text{fpa}}) = C\frac{d\Delta T}{dt} + G\Delta T. \quad (1)$$

Unlike photonic sensors, a microbolometer pixel is *always* exposed to the scene's radiant flux resulting in the so-called *thermal inertia* that prevents abrupt temperature changes in the sensor. Thermal inertia produces a characteristic motion blur with exponentially decaying point spread function that varies spatially [15, 16]. Assuming the incident flux changes from Φ_1 to Φ_2 in a step manner, we can model the change in temperature of the pixel as [6]

$$\Delta T = \frac{\alpha\Phi_1}{G}e^{-\frac{t}{\tau}} + \frac{\alpha\Phi_2}{G}\left(1 - e^{-\frac{t}{\tau}}\right), \quad (2)$$

where α is the conversion efficiency of the microbolometer and $\tau = \frac{C}{G}$ is the time constant of the microbolometer and is a measure of thermal inertia of the pixel. This change in temperature manifests as change in resistance of the microbolometer ΔR

$$\Delta R = \beta R_{\text{avg}}\Delta T, \quad (3)$$

where β is the temperature coefficient of the pixel, and R_{avg} is the average resistance of the microbolometer over the measurement duration. Assuming the pixel reaches steady state within the integration time,

$$\Delta R = \frac{\beta R_{\text{avg}}\alpha(\Phi_2 - \Phi_{\text{fpa}})}{G}. \quad (4)$$

If I be the current flowing through the microbolometer

$$\Delta V = I\Delta R = \frac{\beta R_{\text{avg}}\alpha(\Phi_2 - \Phi_{\text{fpa}})}{G}$$
$$= \frac{I\alpha\beta R_{\text{avg}}}{G}\Phi_2 - \frac{I\alpha\beta R_{\text{avg}}}{G}\Phi_{\text{fpa}}$$
$$\implies \Delta V \equiv g\Phi_2 + o, \quad (5)$$

where g, o are slope and intercept respectively relating the input radiant flux to output voltage. Extending the analysis to all pixels in the sensor for time instance t

$$\Delta V(u,v,t) = g(u,v,t)\Phi_2(u,v,t) + o(u,v,t). \quad (6)$$

Incorporating readout noise in the equation we obtain

$$y(u,v,t) = g(u,v,t)x(u,v,t) + o(u,v,t) + n(u,v,t), \quad (7)$$
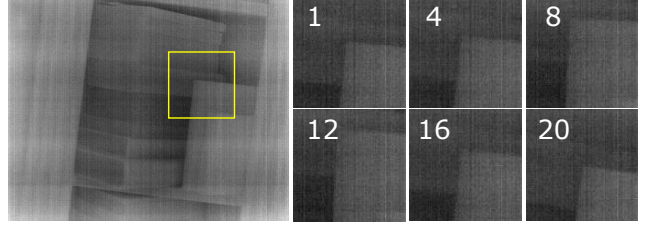

Figure 3: **Effect of camera jitter.** Camera motion affects only the radiant flux entering the camera and not the non-uniformities.

where u, v are camera pixel coordinates, y is the digital output of the camera, and x is scene's radiance that we wish to estimate. Since microbolometers require a small bias current to operate, the temperature changes within the housing, leading to non-uniformities in gain and offset.

Factors affecting offset. Prior works largely assumed that the offset terms in $o(u,v)$ are independent identically distributed Gaussian random variables. However, practical systems have offset contributions from sources that are highly structured, such as internal heating of the camera's housing, or reflections off of the optical subsystem. It is possible to correct for the offset by periodically capturing image of an external black body – however this approach is not always feasible. We instead model the offset term as a spatially smoothly varying signal which can be estimated along with gain and scene's radiant flux.

4. Image Enhancement via Camera Motion

Each frame captured by microbolometer camera is corrupted by an offset term and gain term that is specific to each camera and its operating temperature, which results in an ill-posed system of equations. These gain and offset terms are intrinsic properties of the camera and change slowly over time. This implies, if we were to capture multiple images of the scene over a short interval, any camera motion affects only the scene's radiant flux and not the non-uniformities due to the sensor. This is visualized in Fig. 3 where a sequence of twenty images was captured over a short period. Evidently, the non-uniformities do not change over the duration of twenty frames. This inherent separation motivates our approach — while jitter in the camera is undesirable in visible cameras, it is highly advantageous in the microbolometer camera. As a simple example, consider a capture of 100 images of a scene with and without camera jitter, shown in Fig. 4. A simple averaging does not remove the non-uniformities in the absence of camera jitter. However if we capture with camera jitter and then register all frames to the first reference frame and average, we obtain a relatively noise-free image. DeepIR and averaging rely on camera jitter to recover the scene's radiant flux; however DeepIR requires far fewer images due to a combination of

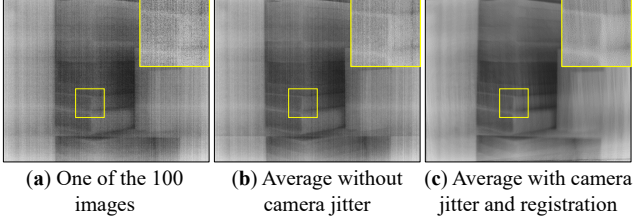| (**a**) One of the 100 images | (**b**) Average without camera jitter | (**c**) Average with camera jitter and registration |

Figure 4: **Advantages of camera jitter.** Small camera motion, while undesirable in visible cameras, helps reduce the effect of slowly changing gain and offset in thermal images.

device physics and concise image representation.

4.1. Modeling multiframe capture

To regularize the inverse problem, we model the gain and offset terms to be constant for the duration of L frames. Further, we assume that every frame can be represented as a geometric transformation of the first frame, which includes rigid, affine, or perspective transforms. The overall model

$$y(u, v, t_k) = g(u, v)(x(f_k(u), h_k(u)) + o(u, v)) \cdots \\ \cdots + n(u, v, t_k), \qquad (8)$$

where f_k, h_k are functions relating pixels in k^{th} frame to first frame. Vectorizing all representations, we obtain

$$\mathbf{y}_k = \mathbf{g} \odot (M_k \mathbf{x}_0 + \mathbf{o}) + \mathbf{n}, \qquad (9)$$

where $\odot$ is element-wise multiplication, M_k is the linear operator to perform the geometric transformation, $\mathbf{g}$ is the gain vector, $\mathbf{x}_0$ is the noise-free image, and $\mathbf{o}$ is the offset term. Our goal is to recover the image $\mathbf{x}_0$. For L frames and N pixels per frame, we have N parameters each from the gain, offset, and latent images, and 8 parameters from the geometric transformation assuming a generalized perspective transformation. Overall, we have LN equations and $3N + 8(L - 1)$ unknowns.

4.2. How much should we jitter the camera?

The amount of jitter needed to accurately recover the scene's radiant flux is highly dependent on the nature of the non-uniformities. Intuitively, the more correlated the spatial non-uniformities, the more the camera needs to jitter. To understand the reason, consider the sequence of L images, $Y_k(u, v) = G(u, v)X_0(u_k, v_k)$, where we assume that the offset is zero. Let us assume that each image is registered back to the reference frame giving us

$$\widehat{Y}_k(u, v) = G(\widehat{u}_k, \widehat{v}_k)X_0(u, v), \qquad (10)$$

where $G(\widehat{u}_k, \widehat{v}_k)$ is the resultant gain after registering $X_k(u, v)$. Then averaging the frame yeilds

$$\widehat{X}_0 = \frac{1}{L} \sum_{k=1}^{L} \widehat{Y}_k(u, v) = \frac{1}{L} X_0(u, v) \sum_{k=1}^{L} G(\widehat{u}_k, \widehat{v}_k). \quad (11)$$

The variance of estimate at pixel (u, v) is then

$$\sigma_{u,v}^2 = \text{Var}\left(\frac{1}{L} X_0(u, v) \sum_{k=1}^{L} G(\widehat{u}_k, \widehat{v}_k) \right)$$

$$= \frac{1}{L^2} X_0^2(u, v) \left(\sum_{k=1}^{L} \text{Var}(g_k) + 2 \sum_{p=1}^{L} \sum_{q=1}^{L} \text{Cov}(g_p g_q) \right). \qquad (12)$$

Equation (12) states that the variance of estimate depends on the autocorrelation function of the gain. Assuming the autocorrelation function monotonically decreases with distance, it is intuitive to see why a more correlated gain requires larger jitter.

In practice, it is difficult to estimate the autocorrelation of the gain as it is a complex function of temperature of operation, and the electronic circuitry. To obtain an empirical estimate of the amount of jitter needed, we imaged a flat black body with the low resolution FLIR lepton, and the medium resolution FLIR Boson cameras. We then computed spatial autocorrelation by cropping random patches and computing cross correlation within a neighborhood of 50 pixels on all sides. Figure 5 shows the captured image, and the temporal and spatial autocorrelation functions for both cameras. We make three observations here. First, the temporal autocorrelation gracefully reduces from 1.0 to 0.6 over 500 frames, with value being greater than 0.8 for up to 20 frames. This implies we can assume approximately constant non-uniformities for up to 20 frames for both cameras. Second, the spatial autocorrelation function is dominated along the horizontal and vertical axes — this is expected since the microbolometer cameras are equipped with a rolling shutter readout circuitry [17]. An immediate implication of this observation is that we cannot achieve noise reduction by just horizontal or vertical shifts, we need a combination of the two. Third, the Lepton camera has an autocorrelation greater than 0.1 over a shift of 3 pixels, and the Boson camera over 5 pixels on either sides. Hence we require non-axial shifts of 3, and 5 pixels, respectively, for the two cameras to ensure high quality reconstruction with a small number of images.

5. Regularizing Physics with Deep Networks

In an ideal scenario, we can estimate the non-uniformities, motion parameters, and the scene's radiant flux from as few as 4 frames. However, due to both signal and readout noises, the inversion is often unstable. Hardie
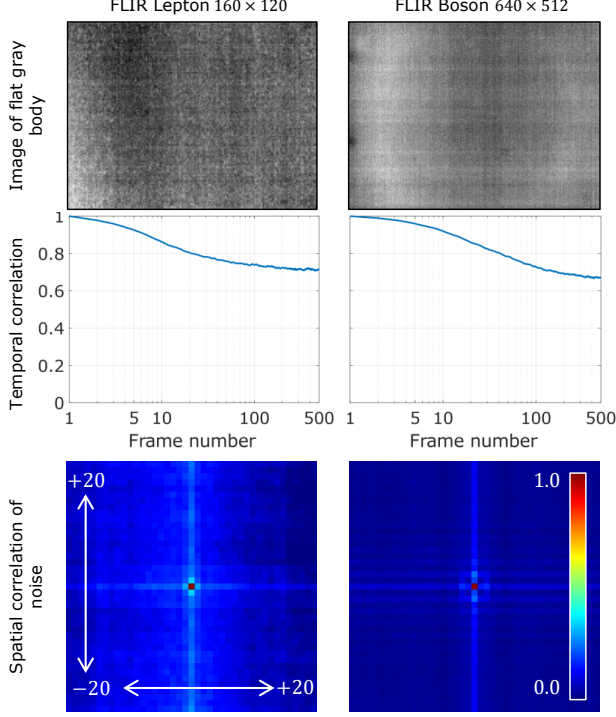
Figure 5: **Statistics of non-uniformities.** The non uniformities associated with thermal cameras have spatial and temporal correlations, which allows us to choose the minimum amount of jitter, as well as the maximum number of frames that are needed to obtain a high quality estimate of the scene's radiant flux.

et al. [7] approached this problem by assuming that the images to be registered, and posing (9) as a least squares problem with 20 input frames. This may not be feasible, as the camera temperature, or the scene may change within that duration. Moreover, in the presence of severe noise, obtaining a reliable registration is difficult. A simple extension is to then jointly optimize for the unknown registration

$$\min_{\mathbf{g},\mathbf{o},M_k,\mathbf{x}_0} \sum_{k=1}^{N} \|\mathbf{x}_k - \mathbf{g} \odot M_k \mathbf{x}_0 - \mathbf{o}\|^2 + TV(\mathbf{x}_0), \quad (13)$$

where $TV(\cdot)$ is the 2D total variation norm, acting as a regularizer. However, the approach fails to converge in the absence of a good initial registration.

This inverse problem can be made tractable if we have a concise representation for the scene's radiant flux. For images in the visible domain, there are several compelling ways of concisely representing images, including analytical signal models [18], or learned representations [19, 20]. Learned models are better tailored to the statistics of real world images and hence have been the choice for image representation. In the presence of a very large pool of data (including noisy and noise-free pairs), it is possible to learn good data-driven models for thermal imaging. However,

due to device-specific noise statistics of each microbolometer camera, such a data-driven approach may not be practical.

5.1. Deep network as a regularizer

Recent works by Ulyanov et. al. [11] on deep image prior have shown that the inductive bias of a convolutional neural networks act as concise priors for images. Specifically, given a fixed input (commonly random noise) $\mathbf{n}$ to a neural network $\mathcal{N}$, deep image prior seeks to solve the following optimization problem,

$$\min_{\mathcal{N}} \|\mathbf{x} - \mathcal{N}(\mathbf{n})\|^2 + \lambda \mathcal{R}(\mathbf{x}), \quad (14)$$

where $\mathbf{x}$ is the signal of interest, and $\mathcal{R}$ is a regularizer specific to signal domain, such as the total variation norm for images. We observe here that the weights of the network $\mathcal{N}$ are learned only with *an* instance of the signal, and not on a pool of data. Such a representation can then used to regularize a range of inverse problems including denoising, super resolution, and inpainting.

5.2. Combining physics and neural representations

Armed with our insights into the sensor physics and ability to concisely regularize images, we now explain how we can efficiently solve for the sensor and scene parameters. We model the scene's radiant flux as the output of a neural network, specifically $\mathbf{i}_0 = \mathcal{N}(\mathbf{p})$, where $\mathcal{N}$ is a convolutional neural network, and $\mathbf{p}$ is a fixed (possibly random) input. We then solve the following optimization problem

$$\min_{\mathbf{g},\mathbf{o},M_k,\mathcal{N}} \|\mathbf{x}_k - \mathbf{g} \odot M_k \mathcal{N}(\mathbf{p}) - \mathbf{o}\|^2 \quad (15)$$

This approach not only preserves the image formation, as well as sensor specific noise characteristics, but also incorporates a concise, non linear representation for image, that has been shown to produce promising results. By optimizing the registration, gain, and the latent image simultaneously, DeepIR accurately estimates the parameters of the system and scene. We make no further assumptions about the structure of the scene, or the non-uniform noise, implying that it works well for cameras with or without shutter-based NUC. Since our approach is modular by separating out the device physics, and the choice of image representation, we can replace the deep image prior with other neural representations such as the deep decoder [21], or implicit neural representations [22, 23]. We show all our experiments with the deep image prior, but extensions with other representations is straightforward.

Solving linear inverse problems. DeepIR can be applied to several linear inverse problems of the form

$$\widehat{\mathbf{r}}_k = \Phi_k \mathbf{r}_k \implies \mathbf{x}_k = \mathbf{g} \odot (\Psi_k \mathbf{r}_k) + \mathbf{o} \quad (16)$$

where Ψ_k is a linear operator. This approach enables us to solve several problems including NUC, super resolution, deconvolution, and optical flow estimation. As an exmaple, we showcase super resolution using DeepIR, where Ψ_k is a downsampling operator.

6. Experiments

We now demonstrate the effectiveness of DeepIR over a diverse set of simulated and real experiments.

6.1. Simulations

We first focus on denoising images. To quantitatively compare various approaches, we took a clean thermal image and added fixed pattern noise, emulating effects due to the readout circuitry specific to microbolometer cameras [17]. Each image was formed in the following manner,

$$Y_k = \mathcal{P}(GS_{\Delta x_k, \Delta y_k}(R_{\theta_k}(X_0))), \qquad (17)$$

where $\mathcal{P}$ is poisson noise operator, G is a random, columnwise gain, S is the shift operator, and R is the rotation operator. We then evaluated the following approaches:

- **Temporal averaging.** We compensate for camera motion by registering the N images and then average them to obtain the latent image.

- **Hardie et al. [8].** We solve the optimization problem in eq. (13) with TV prior on image.

- **He et al. [9].** We used the deep learning-based approach proposed in [9]. For multi-frame comparison, we denoised each individual frame, and then registered and averaged them.

- **DeepIR.** We solve the optimization in (15).

We obtained initial registration between images using the pyramidal registration scheme proposed in [24]. We used the same deep network architecture as proposed in [11] for DeepIR. For all optimization problems, we used PyTorch [25]. Further details about training are provided in the supplementary document.

Non-Uniformity Correction. Figure 6 shows a visualization of various approaches for varying number of images. We observe that incorporating sensor model into the optimization problem enables a significant increase in accuracy, as can be seen in the results of [8], and DeepIR. The advantage of combining sensor model and deep networks is evident in the results of DeepIR, where the reconstruction is significantly better even with as few as three images.

Super Resolution. Figure 7 shows super resolution results on a simulated scene. We simulated a capture of 16 images, downsampled by $4\times$, and then recovered using various approaches. We compared against a modified version
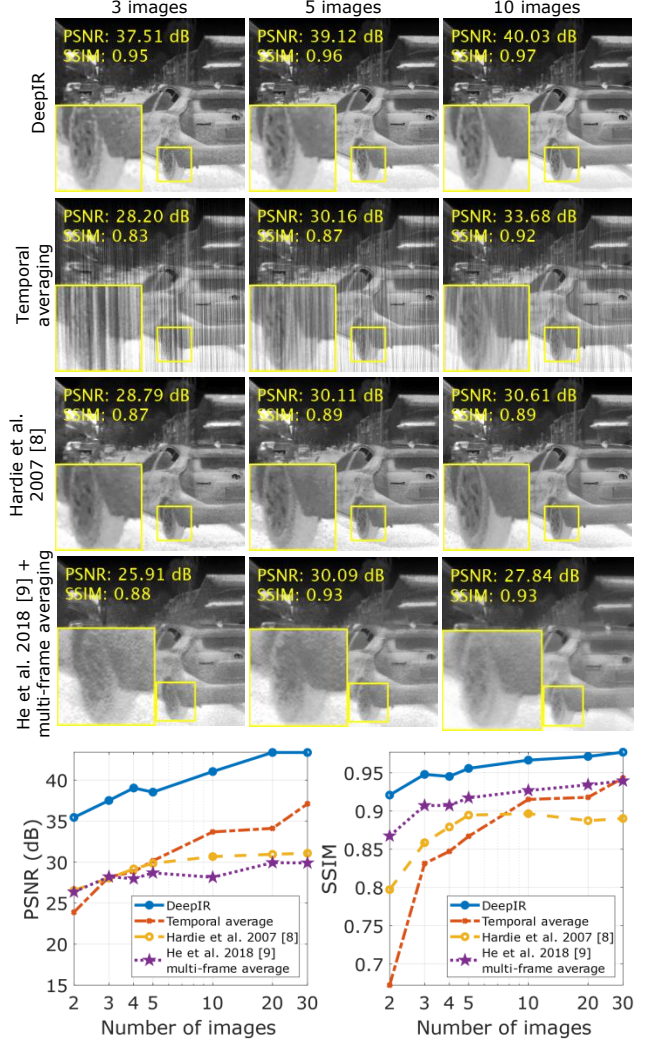

Figure 6: **Performance vs. number of images.** Across the board, DeepIR outperforms other techniques for denoising and NUC.

of Hardie et al. [8] where we added a TV prior to regularize the problem. We also compared with a single image super resolution (SISR) approach trained on visible images [26]. Specifically, we first denoised with He et al.'s [9] approach, and then used this as input to the SISR network.

6.2. Hardware experiments

We captured images with the FLIR Boson 640 to demonstrate NUC over a wide range of scenes, and with FLIR Lepton 3.5 to demonstrate super resolution. The Boson camera captured images at 60 frames per second (fps), and the Lepton captured images at 8.8 fps. In all our experiments, we kept the scene static, and only perturbed the camera, creating a jitter in the images. The Boson camera had an in built mechanical shutter that periodically closes to perform NUC. We performed denoising with and without the shutter-based flat field correction, in order to demonstrate the efficacy of

(a) Ground truth **(b)** DeepIR

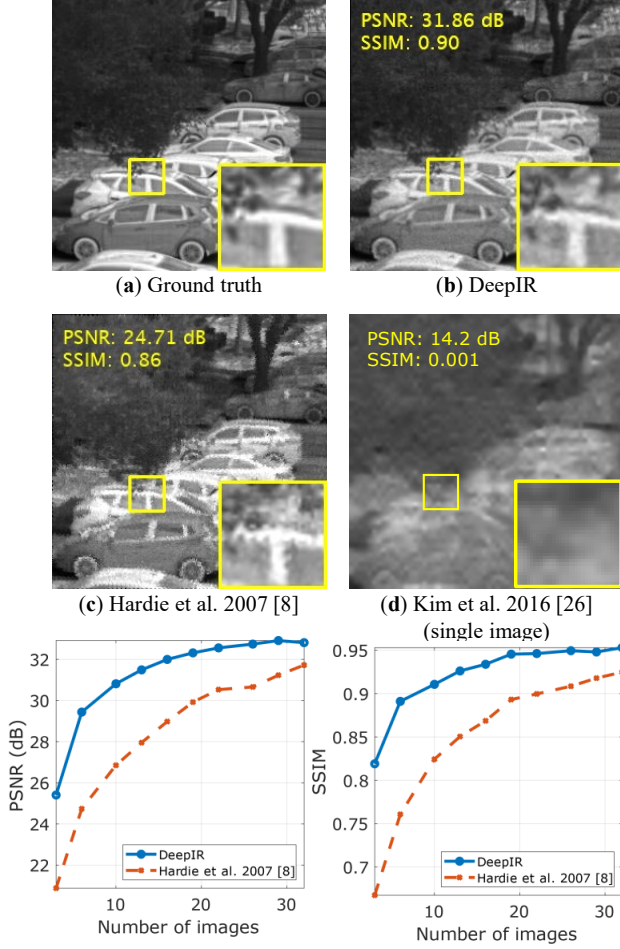**(c)** Hardie et al. 2007 [8] **(d)** Kim et al. 2016 [26]
(single image)

Figure 7: **Simulations for super resolution.** We simulated a capture of 16 images and performed a $4\times$ super resolution with various approaches. DeepIR outperforms competing techniques including single image super resolution [26].

DeepIR across varying hardware settings.

Non-uniformity correction. Figure 8 shows the results of non uniformity correction (NUC) with various approaches. We note that He et al. [9] relied on a large pool of data to learn to specifically remove column-wise NUC. To keep the comparison fair, we applied this approach to multiple images, registered and averaged them. We observe that DeepIR outperforms [8], and is comparable to the quality of [9] with multiple frames averaged.

Figure 9 shows results with Boson's shutter-based flat field correction applied once during the start of the camera, which largely removed the stripes pattern. Since the technique in [9] was not trained for such patterns, the approach did not denoise the image. DeepIR outputs a visibly cleaner image, and estimates a gain that is independent of the scene's geometry.

(a) Frame 1 **(b)** Frame 5

(c) DeepIR **(d)** Hardie et al. 2007 [8]

(e) He et al. 2018 [9] **(f)** He et al. 2018 [9] +
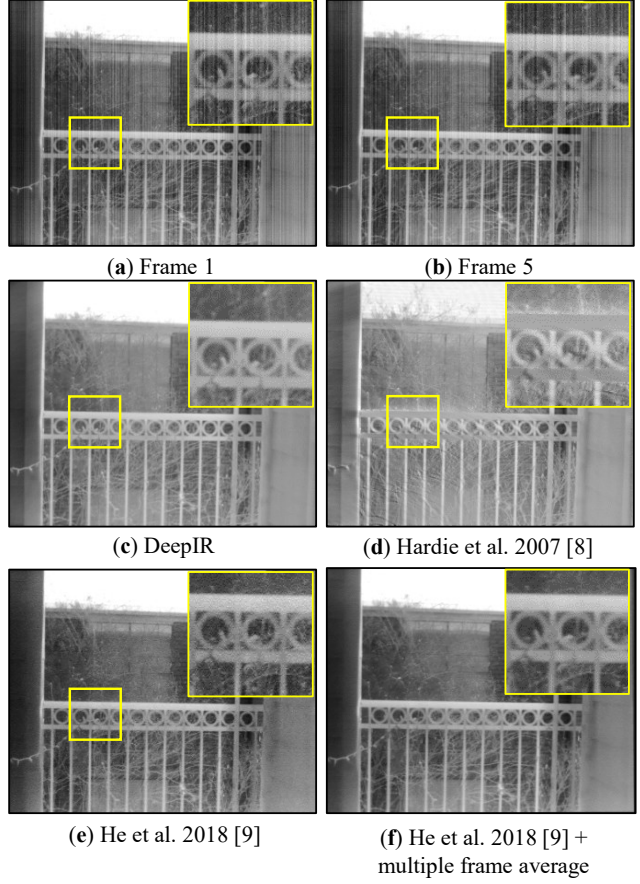multiple frame average

Figure 8: **Non-uniformity correction without shutter-based compensation.** DeepIR performs comparably to supervised techniques [9] on cameras without built-in shutter-based FFC.

Super resolution. The FLIR Lepton camera is a low cost but low resolution thermal imager. In order to increase the resolution by $4\times$, we employed DeepIR framework with 16 images. Notice how the low resolution image contains no significant information about the various keys, but the super resolved image has the keys distinctly separated. DeepIR hence allows us to convert low-cost thermal cameras to high resolution cameras.

Suppressing reflections with a polarizer. Surfaces such as polished metals and glass are strong reflectors in thermal wavelengths, causing interference. Since the reflected polarizataion is predominantly orthogonal to the surface, we can utilize a polarizer to remove its effect. However, the presence of additional optics in front of thermal cameras causes a narcissus effect. Figure 11 visualizes the effect of a polarizer placed in front of our Boson camera. This is a compelling example for how the offset term can be highly structured, and hence biased. We applied DeepIR on the inputs with polarizer to remove the offset term, resulting in an image that was as sharp as the image without a polarizer.
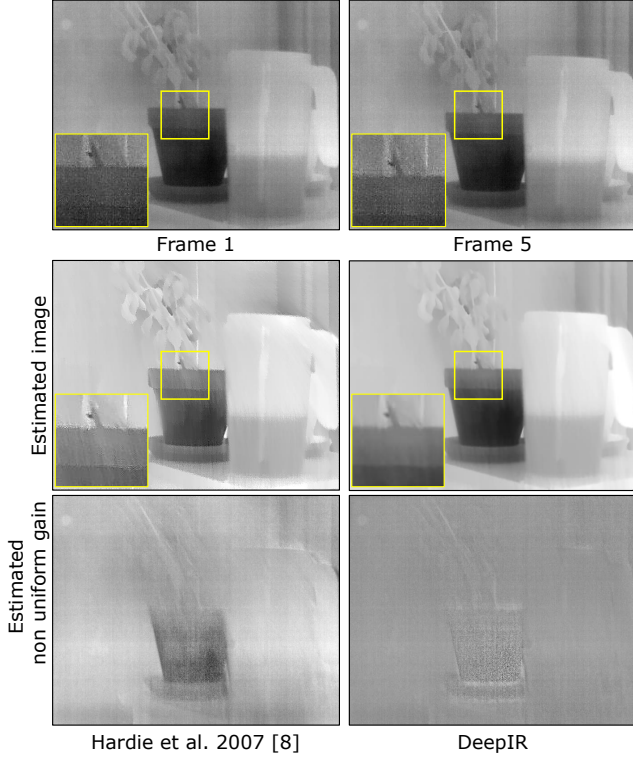
Figure 9: **Non-uniformity correction with shutter-based compensation.** Even with a shutter-based FFC, scenes with very low temperature variations tend to have residual non uniformities. DeepIR can denoise well even under such conditions.
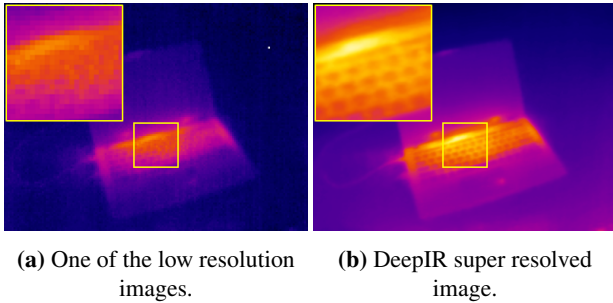

(a) One of the low resolution images.   **(b)** DeepIR super resolved image.

Figure 10: **Super resolution with Lepton camera.** We performed a $4\times$ super resolution with 16 low resolution images captured with a FLIR Lepton to a resolution of 640×480. Images are in false color to show the details. Notice the clearly visible gap in keys.

7. Conclusion

We have developed a general framework called DeepIR for enhancing images in the thermal domain. We achieved this by noting that camera motion, which is usually a hinderance, can be exploited to our advantage to separate a sequence of images into the scene-dependent radiant flux, and a slowly changing scene-independent non-uniformity. DeepIR combines the physics of microbolometer sensors, with powerful regularization capabilities by neural network-
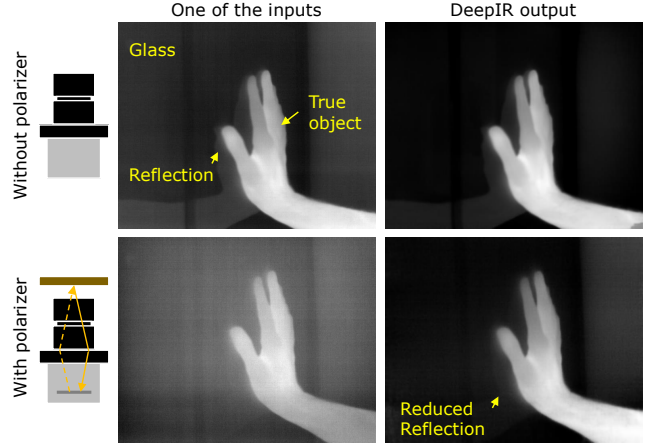

Figure 11: **NUC with external optics.** Addition of optics such as polarizer causes narcissus effect, which increases the image offset. DeepIR is capable of NUC with quality comparable to imaging without polarizer.

based representations. DeepIR relies on the key observation that jittering a camera, while unwanted in visible domain, is highly desirable in the thermal domain as it allows an accurate separation of the sensor-specific non-uniformities from the scene's radiant flux. We showed compelling results on NUC, super resolution, and correction of narcissus effect with external optics. Our framework can be applied to several other tasks in thermal imaging including accurate temperature estimation, motion deblurring, and optical flow.

Current limitations. Since our approach relies on neural networks, the optimization process requires several minutes of GPU computations – resources which preclude a video-rate processing. Future directions may look into speeding up algorithm with implementation improvements, or by using light-weight networks.

Hardware for jittering. While our approach involved manually jittering the camera, it is possible to build hardware systems that can be jittered internally. Some examples include an electromagnetic stage that can precisely control the sensor position within the camera, which is already used in cellphone cameras for image stabilization. Another approach would be to rotate a thick Germanium window in front of the camera that would then produce shifts in various directions. This was the basis of the jitter camera [27] and is another promising direction.

8. Acknowledgements

A. Learning Details

All the results in the paper were regularized with a deep image prior based regularize. Our goal was to demonstrate the advantages of combining physics and deep networks, and hence our network architecture was an unmodified version of the architecture utilized in the original paper [11]. Specifically, we used a convolutional network with skip connections shown in Fig. 12. We note that alternate networks are possibly and potentially capable of giving better results but was not the focus of our paper.

Optimization details. As mentioned in the paper, we jointly optimized the parameters of the neural network, 6 parameters for each of the N affine matrices, $H \times W$ dimensional gain and offset terms. The input to the neural network was a $H \times W \times 8$ shaped noise that was *not* optimized along with other parameters. We found that random initialization for affine matrices sufficed – however to accelerate convergence we first registered the images to the first image using a pyramidal registration algorithm [24].

Details about super resolution. The image formation model relating the low resolution image $\mathbf{x}_k$ and high resolution image $\mathbf{x}_{\text{HR}}$ is,

$$\mathbf{x}_k = \mathbf{g} \odot (D M_k \mathbf{x}_{\text{HR}}) + \mathbf{o}, \tag{18}$$

where D is the downsampling operator, and M_k is the transformation matrix. To prevent aliasing artifacts endemic to downsampling, we chose D as the following operation,

$$X_{\text{LR}}(u,v) = \frac{1}{Q^2} \sum_{p=1}^{Q} \sum_{q=1}^{Q} X_{\text{HR}}(u+p, v+q) \tag{19}$$

for downsampling by a factor of Q.

Learning parameters. We set the learning rate to 10^{-3} and trained for a total of $2,000$ iterations. For non-uniform correction, there was no penalty for optimizing beyond $2,000$ iterations. However increasing the number of iterations proved to be detrimental for super resolution by producing checker-like artifacts in the final reconstruction This is expected, as deep image prior tends to overfit to noise if run for too many iterations.

Our loss function consisted of MSE loss between predicted image $\mathbf{g} \odot M_k (\mathbf{x}_o + \mathbf{o})$ and the ground truth $\mathbf{x}_k$, a 2D total variation (TV) prior on the latent image, and a TV loss on the offset term. The motivation behind the TV loss for the offset is due to it arising from reflections off of optics which tend to be spatially smooth. We found this to be an effective strategy in separating the gain and offset terms. We set the weight of the TV loss on the latent image to be 10^{-5}, and the weight of the TV loss on the offset term to be 10. We used a batch size equal to the number of input images. The model was trained a system with Nvidia RTX
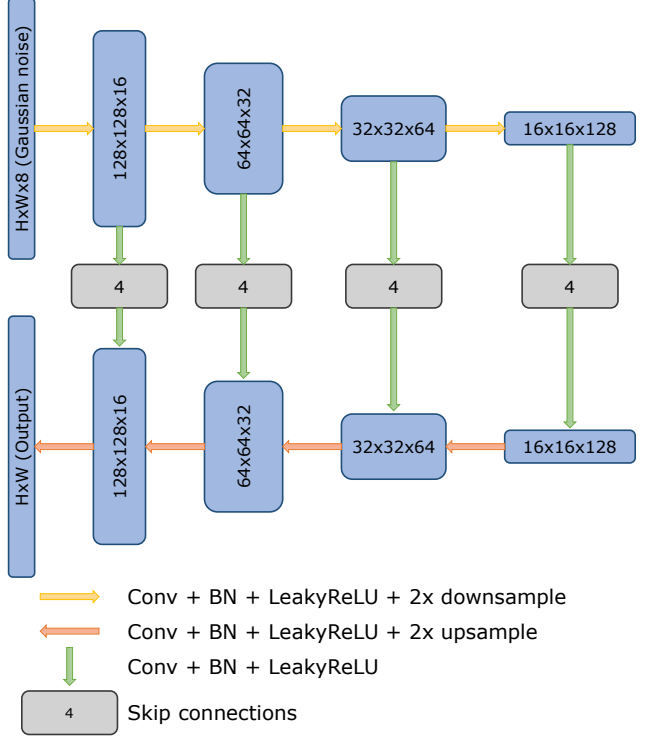

Figure 12: **Network architecture.** We used the default network architecture proposed in [11] for super resolution.

2080 GPU with 8GB memory along with 48GB RAM. The optimization was implemented with the `pytorch` framework [25]. The code ran for 10 minutes on our computer for five images of size 640×512 for a total of $2,000$ iterations. We will release our optimization code to the public for further research in this direction.

B. Real Results

We demonstrate some more results and provide sensitivity to parameters.

Hardware details. We used the FLIR Boson camera with 640×512 spatial resolution capturing images at 60 frames per second (fps), and the FLIR Lepton camera with 160×120 spatial resolution capturing images at 9 fps. We used the `flirpy` [28] package to control the cameras which allowed us to disable periodic NUC and capture images at full frame rate of the individual cameras. The Boson camera was equipped with inbuilt flat field correction (FFC), supplementary correction for lens reflections, and temporal noise reduction. We showed results with and without FFC in the main paper. In all cases, we disabled temporal noise reduction, as we found that enabling it produced ghosting artifacts.
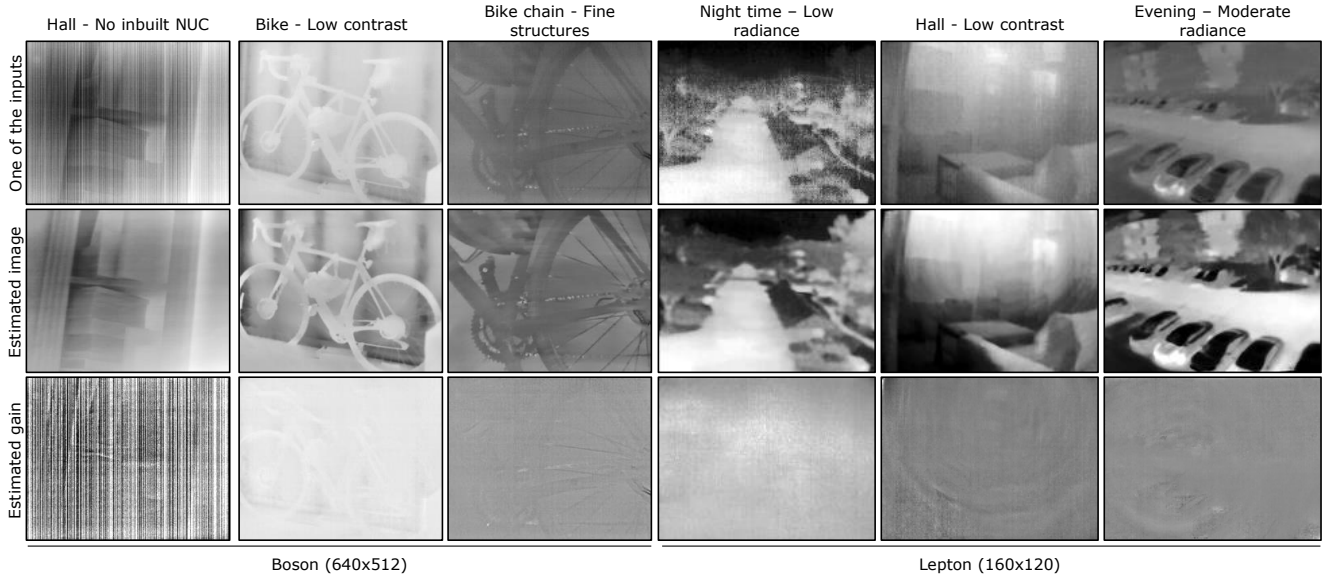
Non-uniformity correction. We showed NUC results on

Figure 13: **NUC on diverse scenes.** Our approach is capable of non-uniformity correction for a wide variety of noise levels and scene complexities.
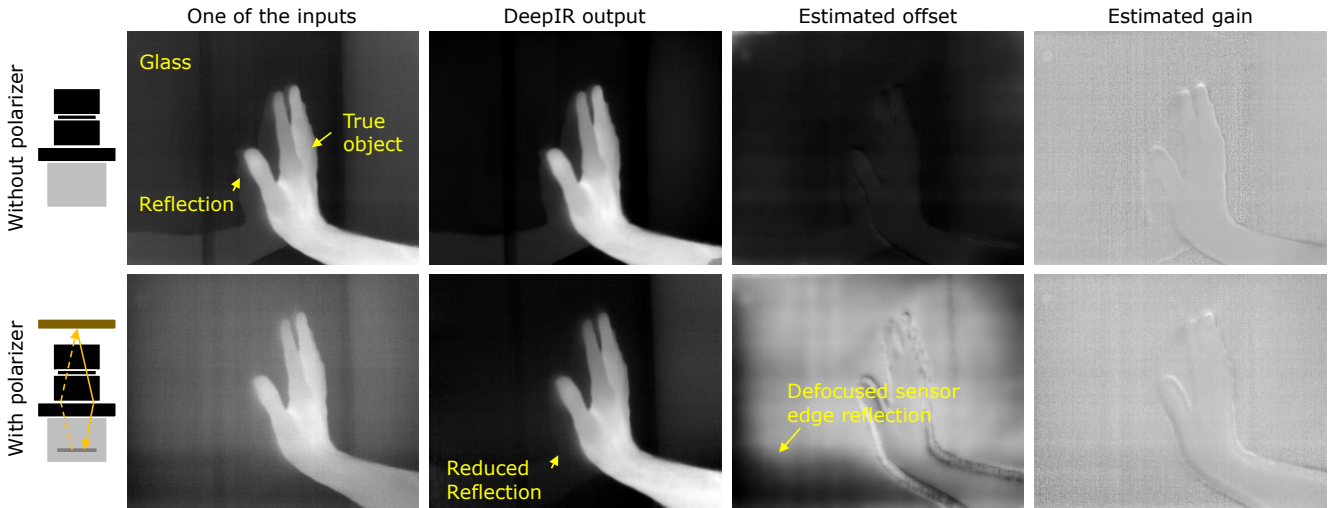

Figure 14: **Suppressing narcissus effect.** Since we model both gain and offset terms, DeepIR is capable of removing narcissus effects due to external optics like polarizers.

some scenes with the Boson camera in the main paper. We next demonstrate some more experiments to underline the advantages of DeepIR. Figure 13 shows the non-uniformity correction with the various scenes at varying levels of scene complexity. All experiments included recovery with five images. We found the offset to be nearly zero and hence did not visualize it. DeepIR performs promisingly in low contrast conditions, absence of inbuilt NUC, low and low radiance levels.

Suppressing narcissus effect. Figure 14 shows the images with and without polarizer. Since we model both gain and

offset, we were able to suppress the narcissus effect arising out of back reflections from the polarizer. Notice the defocused edge that is visible in the estimated offset in the image captured with a polarizer. The edge artifacts looking like the hard were due to minor motion between frames, and can be corrected with a more accurate model of transformation such as optical flow.

References

[1] FLIR thermal dataset for algorithm training. https://www.flir.com/oem/adas/

`adas-dataset-form/`, 2021. 1

[2] Marina Ivašić-Kos, Mate Krišto, and Miran Pobar. Human detection in thermal imaging using yolo. In *Intl. Conf. Computer and Technology Applications*, 2019. 1

[3] Mate Krišto, Marina Ivasic-Kos, and Miran Pobar. Thermal object detection in difficult weather conditions using YOLO. *IEEE Access*, 8:125459–125476, 2020.

[4] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *IEEE Comp. Vision and Pattern Recognition (CVPR)*, pages 1037–1045, 2015. 1

[5] Olivier Janssens, Rik Van de Walle, Mia Loccufier, and Sofie Van Hoecke. Deep learning for infrared thermal image based machine health monitoring. *IEEE Trans. Mechatronics*, 23(1):151–159, 2017. 1

[6] Michael Vollmer and Klaus-Peter Möllmann. *Infrared Thermal Imaging: Fundamentals, Research and Applications*. John Wiley & Sons, 2017. 1, 3

[7] Russell C Hardie, Majeed M Hayat, Earnest Armstrong, and Brian Yasuda. Scene-based nonuniformity correction with video sequences and registration. *Appl. Optics*, 39(8):1241–1250, 2000. 1, 2, 5

[8] Russell C Hardie and Douglas R Droege. A map estimator for simultaneous superresolution and detector nonunifomity correction. *J. Adv. in Signal Processing*, 2007:1–11, 2007. 1, 2, 6, 7

[9] Zewei He, Yanpeng Cao, Yafei Dong, Jiangxin Yang, Yanlong Cao, and Christel-Löic Tisse. Single-image-based nonuniformity correction of uncooled long-wave infrared detectors: A deep-learning approach. *Appl. Optics*, 57(18):D155–D164, 2018. 1, 2, 6, 7

[10] Rafael E Rivadeneira, Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Thermal image superresolution through deep convolutional neural network. In *Intl. Conf. Image Analysis and Recognition*, 2019. 1

[11] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *IEEE Comp. Vision and Pattern Recognition (CVPR)*, 2018. 2, 5, 6, 9

[12] Robert Olbrycht and Bogusław Więcek. New approach to thermal drift correction in microbolometer thermal cameras. *Quantitative InfraRed Thermography Journal*, 12(2):184–195, 2015. 2

[13] Chengwei Liu, Xiubao Sui, Guohua Gu, and Qian Chen. Shutterless non-uniformity correction for the long-term stability of an uncooled long-wave infrared camera. *Measurement Science and Technology*, 29(2):025402, 2018. 2

[14] Alejandro Wolf, Jorge E Pezoa, and Miguel Figueroa. Modeling and compensating temperature-dependent non-uniformity noise in IR microbolometer cameras. *Sensors*, 16(7):1121, 2016. 2

[15] Manikandasriram Srinivasan Ramanagopal, Zixu Zhang, Ram Vasudevan, and Matthew Johnson-Roberson. Pixel-wise motion deblurring of thermal videos. *arXiv preprint arXiv:2006.04973*, 2020. 3

[16] Beata Oswald-Tranta, Mario Sorger, and Paul O'Leary. Motion deblurring of infrared images from a microbolometer camera. *Infrared Physics and Technology*, 53(4):274–279, 2010. 3

[17] Musaed Alhussein and Syed Irtaza Haider. Simulation and analysis of uncooled microbolometer for serial readout architecture. *J. Sensors*, 2016, 2016. 4, 6

[18] S Grace Chang, Bin Yu, and Martin Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans. Image Processing*, 9(9):1532–1546, 2000. 5

[19] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Processing*, 54(11):4311–4322, 2006. 5

[20] JH Rick Chang, Chun-Liang Li, Barnabas Poczos, BVK Vijaya Kumar, and Aswin C. Sankaranarayanan. One network to solve them all–Solving linear inverse problems using deep projection models. In *IEEE Comp. Vision and Pattern Recognition (CVPR)*, 2017. 5

[21] Reinhard Heckel and Paul Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. *Intl. Conf. Learning Representations*, 2018. 5

[22] Vincent Sitzmann, Julien NP Martel, Alexander W Bergman, David B Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *arXiv preprint arXiv:2006.09661*, 2020. 5

[23] Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *arXiv preprint arXiv:2006.10739*, 2020. 5

[24] Philippe Thevenaz, Urs E Ruttimann, and Michael Unser. A pyramid approach to subpixel registration based on intensity. *IEEE Trans. Image Processing*, 7(1):27–41, 1998. 6, 9

[25] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. 2019. 6, 9

[26] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Comp. Vision and Pattern Recognition (CVPR)*, 2016. 6, 7

[27] Moshe Ben-Ezra, Assaf Zomet, and Shree K Nayar. Jitter camera: High resolution video from a low resolution detector. In *IEEE Comp. Vision and Pattern Recognition (CVPR)*, 2004. 8

[28] Flirpy. `https://flirpy.readthedocs.io/en/latest/`, 2021. 9